

## DOCUMENT RESUME

ED 135 028

CS 501 612

TITLE Status Report on Speech Research; A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, October 1 - December 31, 1976.

INSTITUTION Haskins Labs., New Haven, Conn.

REPORT NO SR-48 (1976)

PUB DATE 76

NOTE 361p.

EDRS PRICE MF-\$0.83 HC-\$19.41 Plus Postage.

DESCRIPTORS \*Acoustic Phonetics; \*Articulation (Speech); \*Artificial Speech; \*Audition (Physiology); \*Auditory Perception; Beginning Reading; \*Language Research; Phonemes; Phonetics; Psychoacoustics; Spectrograms

## ABSTRACT

This report is one of a regular series on the status and progress of studies on the nature of speech, instrumentation of its investigation, and practical applications. The seventeen papers treat such topics as issues in theory of action; physiological aspects of speech production; role of phonetic structure in linguistic communication; difference limens for formant frequencies in vowel production; vocal tract normalization; speech, the alphabet, and learning to read; role of silence in the perception of stop consonants; perception of implosive transitions in VCV utterances; categorical perception; right-ear advantage for speech heard on the left; left-ear advantage for sounds of rapidly varying resonance frequency; information processing and speech perception; rules for surname pronunciation; and an -s detection algorithm. (AA)

\*\*\*\*\*  
\* Documents acquired by ERIC include many informal unpublished \*  
\* materials not available from other sources. ERIC makes every effort \*  
\* to obtain the best copy available. Nevertheless, items of marginal \*  
\* reproducibility are often encountered and this affects the quality \*  
\* of the microfiche and hardcopy reproductions ERIC makes available \*  
\* via the ERIC Document Reproduction Service (EDRS). EDRS is not \*  
\* responsible for the quality of the original document. Reproductions \*  
\* supplied by EDRS are the best that can be made from the original. \*  
\*\*\*\*\*

U.S. DEPARTMENT OF HEALTH  
EDUCATION & WELFARE  
NATIONAL INSTITUTE OF  
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

SR-48 (1976)

Status Report on  
SPEECH RESEARCH

A Report on  
the Status and Progress of Studies on  
the Nature of Speech, Instrumentation  
for its Investigation, and Practical  
Applications

1 October - 31 December 1976

Haskins Laboratories  
270 Crown Street  
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports).

25 501 612

ED135028



ACKNOWLEDGEMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Dental Research  
Grant DE-01774

National Institute of Child Health and Human Development  
Grant HD-01994

Assistant Chief Medical Director for Research and Development,  
Research Center for Prosthetics, Veterans Administration  
Contract V101(134)P-342

Advanced Research Projects Agency, Information Processing  
Technology Office, under contract with the Office of  
Naval Research, Information Systems Branch  
Contract N00014-76-C-0591

United States Army Electronics Command, Department of Defense  
Contract DAAB03-75-C-0419(L 433)

National Institutes of Child Health and Human Development  
Contract N01-HD-1-2420

National Institutes of Health  
General Research Support Grant RR-5596

*ERRA SHEET ADDED.  
SEE PAGES 362 and 363.*

HASKINS LABORATORIES  
Personnel in Speech Research

Alvin M. Liberman,\* President and Research Director  
Franklin S. Cooper, Associate Research Director  
Patrick W. Nye, Associate Research Director  
Raymond C. Huey, Treasurer  
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson\*  
Thomas Baer  
Peter Bailey<sup>1</sup>  
Fredericka Bell-Berti\*  
Gloria J. Borden\*  
Robert Crowder\*  
James E. Cutting\*  
Donna Erickson  
Frances J. Freeman\*  
Jane H. Gaitenby  
Thomas J. Gay\*  
Katherine S. Harris\*  
Alice Healy\*  
Frances Ingemann\*  
Isabelle Y. Liberman\*  
Leigh Lisker\*  
Ignatius G. Mattingly\*  
Paul Mermelstein  
Seiji Niimi<sup>2</sup>  
Lawrence J. Raphael\*  
Bruno H. Repp  
Philip E. Rubin  
Donald P. Shankweiler\*  
Linda Shockey  
George N. Sholes  
Michael Studdert-Kennedy\*  
Quentin Summerfield<sup>1</sup>  
Michael T. Turvey\*  
Robert Verbrugge\*

Technical and Support Staff

Eric L. Andreasson  
Elizabeth P. Clark  
Harriet Greisser\*  
Donald Hailey  
Terry Halwes  
Elly Knight\*  
Sabina D. Koroluk  
Agnes McKeon  
Terry F. Montlick  
Nancy R. O'Brien  
Loretta J. Reiss  
William P. Scully  
Richard S. Sharkany  
Leonard Szubowicz\*  
Edward R. Wiley  
David Zeichner

Students\*

|                    |                  |                    |
|--------------------|------------------|--------------------|
| Katherine Benner   | Nieba Jones      | Lee Perrin         |
| Steve Braddon      | Lynn Kerr        | Sandra Prindle     |
| David Dechovitz    | Morey J. Kitzman | Abigail Reilly     |
| Laurel Dent        | Andrea J. Levitt | Robert Remez       |
| Susan Lea Donald   | Roland Mandler   | Helen Simon        |
| F. William Fischer | Leonard Mark     | Emily Tobey        |
| Hollis Fitch       | Nancy McGarr     | Harold Tzeutschler |
| Anne Fowler        | Georgia Nigro    | Michele Werfelman  |
| Carol A. Fowler    | Mary Jo Osberger |                    |

\*Part-time

<sup>1</sup>Visiting from The Queen's University of Belfast, Northern Ireland.

<sup>2</sup>Visiting from University of Tokyo, Japan.

## CONTENTS

### I. Manuscripts and Extended Reports

|                                                                                                                                                             |     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Issues in the Theory of Action -- Michael Turvey, Robert Shaw,<br>and William Mace . . . . .                                                                | 1   |
| Physiological Aspects of Speech Production: Why Study Speech<br>Production? -- Katherine S. Harris . . . . .                                                | 21  |
| Universals in Phonetic Structure and Their Role in Linguistic<br>Communication -- Michael Studdert-Kennedy . . . . .                                        | 43  |
| Difference Limens for Formant-Frequencies for Steady-State and<br>Consonant-Bounded Vowels -- Paul Mermelstein and Hollis Fitch . . . . .                   | 51  |
| Vocal Tract Normalization for /s/ and /ŝ/ -- Janet May . . . . .                                                                                            | 67  |
| Speech, the Alphabet and Teaching to Read -- Isabelle Y. Liberman<br>and Donald Shankweiler . . . . .                                                       | 75  |
| Visual Processing and Short-Term Memory -- Michael Turvey . . . . .                                                                                         | 97  |
| Contrasting Orientations to the Theory of Visual Information-<br>Processing -- Michael Turvey . . . . .                                                     | 153 |
| Evidence for a Special Speech-Perceiving Subsystem in the Human --<br>Alvin M. Liberman and David B. Pisoni . . . . .                                       | 183 |
| Further Observations on the Role of Silence in the Perception of<br>Stop Consonants -- Michael Dorman, Alvin M. Liberman, and Lawrence<br>Raphael . . . . . | 199 |
| Perception of Implosive Transitions in VCV Utterances --<br>Bruno H. Repp . . . . .                                                                         | 209 |
| What Can /w/, /l/, /y/ Tell Us About Categorical Perception? --<br>Lyn Frazier . . . . .                                                                    | 235 |
| Laterality and Localization: A Right-Ear Advantage for Speech<br>Heard on the Left -- Christopher Darwin, Peter Howell, and<br>Susan A. Brady . . . . .     | 257 |
| Left-Ear Advantage for Sounds Characterized by a Rapidly Varying<br>Resonance Frequency -- Mark Blechner . . . . .                                          | 279 |
| An Information-Processing Approach to Speech Perception --<br>James Cutting and David Pisoni . . . . .                                                      | 287 |

|                                                                                                                  |     |
|------------------------------------------------------------------------------------------------------------------|-----|
| Outline of a Surname Pronunciation - Rules for a Reading Machine --<br>Jane Gaitenby and S. Lea Donald . . . . . | 327 |
| Building an -S Detection and Removal Algorithm -- George Sholes . . . . .                                        | 339 |
| II. <u>Publications and Reports</u> . . . . .                                                                    | 345 |
| III. Appendix: DDC and ERIC numbers (SR-21/22-SR-45/46) . . . . .                                                | 347 |

I. MANUSCRIPTS AND EXTENDED REPORTS



Michael T. Turvey†, Robert E. Shaw††, and William Mace†††

ABSTRACT

It is argued that the variables of the effector mechanisms cannot be individually and directly controlled. The programming of coordinated movements is not in terms of individual muscles but in terms of coordinative structures where a coordinative structure is defined as a set of muscles, often spanning many joints, that is constrained to act as a unit. The question of how individual coordinative structures are organized is raised. On examination, it appears that the activities of individual muscles composing a coordinative structure relate in terms of a ratio that is invariant over magnitude changes in these activities. The question of the relationship between transport movements that are in reference to the global invariants such as gravity is also raised. A general analysis of the link between perceiving and acting leads to the hypothesis that there is a precise mathematical relationship between coordinative structures and perceptual variables--that the action/perception system provides sui generis its own control.

The Degrees of Freedom Problem

We may juxtapose two fundamental approaches to designing a machine that acts within and upon an environment (cf. Greene, in press a). In one approach, the variables of the machine are programmed to comply with our wishes: a desired action is achieved by virtue of a single computation that in a single instance specifies all the necessary details, including those needed to immunize the machine against perturbing influences. The other

---

\*This paper will be appear in Attention and Performance VII, ed. by J. Requin. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

† So University of Connecticut, Storrs.

††University of Connecticut, Storrs.

†††Trinity College, Hartford, Connecticut.

Acknowledgement: The authors wish to acknowledge Carol Fowler and Robert Remez for their contributions to the authors' appreciation of the problems approached by this paper.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

approach begins with the construction of autonomous systems designed in part, to preserve the stability of the machine in its intended environment. Given these systems that perform as they wish, we proceed to organize them in such a fashion that the systems, and thus the machine, perform as we wish.

In the first kind of machine, executive procedures are needed to control each variable individually while in the second, they are needed to control only a subset of the total. Let us elaborate: where the total number of variables is quite small, the first design is obviously felicitous, but where that number is large, then it is roughly apparent that the first design will prove to be cumbersome and costly. By contrast, a machine of the second kind would be inelegant where few variables were concerned, since any arbitrary configuration of those variables could not be achieved directly, but only indirectly through the modulation and interaction of the autonomous subsystems. A subtle advantage of the second kind of machine is that, with preformed "partial actions" at its disposal, it may achieve an approximation to a desired configuration of a very large number of variables through the individual control of very few. In brief, our intuition is that where the number of variables is excessive, the second design is the more practical.

There are several additional and very significant distinctions to be drawn between the two kinds of machine, but let us proceed to develop the argument that the second kind is more representative of humans and animals than is the first (Greene, 1972, in press a; Turvey, in press).

A classical point of view in the physiology of movement coordination is that the corticospinal projection contains all the details relating to the spatial and temporal patterning of commands to muscles. The classical view promotes a machine of the first type, since it assumed that a coordination of movements results from a single stage of exact computation. However, it has been argued at some length and by many scholars over the years (for example, Paillard, 1960; Bernstein, 1967) that this conception of movement control is simply untenable. Nevertheless, it is our impression that where matters of coordination are discussed, the classical view is very often unwittingly taken as the backdrop to the discussion. We would do well, therefore, to remind ourselves of the reasons (d'apres Bernstein, 1967) that militate against the classical conception noting in anticipation that they do so by underscoring the indeterminacy of the effector mechanisms and the functional equivocality that exists between the systems that plan and the systems that do.

In the first place, the indeterminacy of the anatomical structures is recognized. The movements at the joints and the permissible motions of the complex biokinematic chains that make up the skeletomuscular hardware of animals comprise an inordinately large number of degrees of freedom. Further, the muscles that on some accounts might be taken as the targets of central control commands, are ambiguous in their roles with regard to joint movement. Consider, by way of example, the upper pectoralis major that inserts proximally in the clavicle and distally in the upper shaft of the humerus. With the arm in an approximately horizontal position, in which the

axis of the humerus is just below the horizontal axis of the shoulder joint, contraction of the pectoralis will adduct the arm in the horizontal plane. However, from an approximately horizontal position, in which the axis of the humerus is slightly above the horizontal axis of the shoulder joint, contraction will adduct the arm in the vertical plane (Wells, 1961). The moral (for a brain as well as for a student of kinesiology) is that a muscle's role cannot be taken for granted; at each phase of a movement, an individual muscle's action is contingent on the muscle's line of pull to the joint's axis of motion.

Cognate with this class of equivocalities is the realization that the role a muscle plays depends not only on the disposition of limb segments, but also on the external force contingencies. Lowering the arm from a horizontal side position against a resistance requires the use of the adductors of the arm, notably, the latissimus dorsi; but in lowering slowly (that is, against gravity), the adductors are palpably soft, for the responsibility of the movement befalls the abductors, the deltoids, that perform their task by lengthening or, as Hubbard (1960) prefers to call it, pliometric contraction.

In the second place, there is the indeterminacy resulting from mechanical sources. Most notable among these is the fact that, depending on the dynamic and static conditions of the limb segment, the same innervational state of a muscle may give rise to a variety of motions of the segment differing in displacement and velocity, and different innervational states may produce identical motions. The lesson here is a simple one: innervational states and movement relate equivocally. One is reminded that the much sought after invariance in electromyographic (EMG) records relating to articulatory gestures in speech production (cf. Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967), was rudely repelled by the finding of ubiquitous variance (MacNeilage, 1970).\*

A closely related source of mechanical indeterminacy is indigenous to multilink kinematic chains of which a whole arm, a whole leg, or the whole body are examples. Quite simply, the movement of any one link will result in a displacement of the links attached, but not without reprisal; whereas the attached or "light" links may be passively carried by the agonist or "heavy" link (in the terminology of Eshkol and Wachman, 1958), their passive motions will induce forces and moments as reactions to the "heavy" links trajectory, and thereby complicate its control. From the perspective of the problem of coordination, multilink biokinematic chains look capricious. Nevertheless, facility with a gross body skill is synonymous with exploiting these reactive consequences to the fullest (Bernstein, 1967).

The two sources of mechanical indeterminacy most evidently go hand-in-hand. Because the links have mass, once impelled, they gather momentum and develop kinetic energy. A given degree of muscle activity acting against a movement may stop it, simply retard it, or even reverse it; the same degree of activity, in concert with the movement, may induce marked acceleration. What follows from a given degree of muscle activity depends on the kinetic conditions of the links. Unfortunately, the significance and ubiquity of

25

this principle is rarely recognized and, more often than not, is simply ignored. In very large part, this negligence seems to be due to the popular assumption that the innervational states of muscles are in phase or concurrent with the movements of biokinematic links (Hubbard, 1960), an assumption that deserves our attention, if only briefly.

The "in-phase" assumption is a most convenient one, because as Hubbard (1960) points out, it permits the luxury of inferring muscle events from movement events. For example, as the elbow flexes, the biceps shortens and the triceps lengthens--from which we might infer that there was continuous graded stimulation contracting the one muscle and relaxing the other.

Fast movements, often referred to as ballistic, are anomalous from the perspective of the "in-phase" assumption. Their control is characteristically "bang-bang" (Arbib, 1972); an initial burst of acceleration as the agonist contracts, an intervening period of inactivity, and then a burst of deceleration as the antagonist acts to degenerate the kinetic energy of the link. In fast movements, muscle activity is simultaneous with only a small portion of the movement. But perhaps the "in-phase" assumption does hold for movements conducted at a more leisurely pace, movements that we might refer to as nonballistic. Hubbard has argued and demonstrated that even here the "in-phase" assumption is found wanting (Hubbard, 1960); as far as he can discern, the basis of slow movements is the same as that of fast movements--that is, discrete bursts of muscle activity that alternately act to accelerate and decelerate the link. The control of slow movements, in this perspective, is characteristically "bang-bang-bang", and so on, and there is considerable support for this characterization (for example, Aizerman and Andreeva, 1968; Chernov, 1968; Litvintsev, 1968). It appears that the slow movement of a link, say, wrist extension or elbow flexion, is the result of pulls by both opposing muscles where each muscle pulls ten times per second (Hubbard, 1960; Aizerman and Andreeva, 1968), first one and then the other.

In the third and final place, we may recognize the indeterminacy that arises by virtue of the physiology. We can relate here only a small part of what is, most obviously, a very lengthy story.

The motor unit, conventionally defined as an alphamotoneuron, together with the bundle of extrafusal muscle fibers that it innervates, may be considered as the functional final common path. Alphamotoneurons have monosynaptic connections with several descending systems, but this is essentially true only for primates, and then only for the more peripheral effectors, such as the muscles of the fingers. Even so, the monosynaptic projections to alphamotoneurons represent but a small part of the total neural projection to these cells, and in addition, monosynaptic projections in and of themselves probably do not bring about motoneuron firing (Evarts, Bizzi, Burke, DeLong, and Thach, 1971). In very large part, the major influences ultimately exerted on motor units occur by way of the segmental interneurons. Supra-, inter-, and intra-segmental influences converge on the segmental interneurons so that their modulation of motoneuron activity is highly flexible. Significantly, the same descending "instruction" might at



different times encounter quite different "states" in the segmental interneurons; its effect, therefore, on a target motorneuron, is open to considerable variation.

The fundamental point to be made is that the segmental apparatus of the spinal cord is an active apparatus that does not passively reproduce supraspinal instructions (Paillard, 1960). On the contrary, it appears that supraspinal and spinal influences relate coalitionally (see Turvey, in press) in the coordination of acts. There is evidence that the state of the segmental apparatus can (among other things) convert a flexion reflex into one of extension (Lisin, Frankstein and Rechtman, 1973), enhance or inhibit contractile states evoked by cortical stimulation (for example, Gellhorn, 1948), and affect the latency of voluntary movement (Gurfinkel and Paltsev, 1965).

This overview of "peripheral indeterminacy" is brief and fragmentary, but it is sufficient to point out that the variables of the effector mechanisms are not individually and directly controllable--that the programming of coordinated movements cannot be in terms of individual muscles. On the more positive side, it permits us to identify, after Bernstein (1967), the basic problems of coordination: the mastery of the abundant degrees of freedom attainable by the motor apparatus. We may state the problem in a slightly different way, one that emphasizes the disparity between "central planners" and "peripheral doers": the degrees of freedom of the systems controlling action are far less than the mechanical degrees of freedom of the systems they control (Kots, Krinsky, Naydin and Shik, 1971).

With reference to the two kinds of machines that we outlined at the beginning, let us ask: What constrains the evolution of an action plan into a particular biokinematic trajectory? It is readily apparent that in the perspective of the first kind of machine the answer is simple, since a plan is seen to constrain completely the kinematic degrees of freedom and, therefore, prescribes unambiguously a particular biokinematic configuration. As we now understand, the modus operandi of this machine (and, therefore, its answer to our question) rests on the assumption of an unequivocal relation between "motor commands" and "motor effects." In that this assumption is false for natural action-systems, the first kind of machine cannot be taken as representative of nature's design and, further, an action plan cannot be said to constrain, except lightly, the kinematic degrees of freedom. The infeasibility of prescribing muscular and mechanical details suggests that the planning and executing of an act is to be understood as a flow of control (rather than a chain of commands) that selectively and progressively freezes-out kinematic degrees of freedom.

The second kind of machine has a style of coordination that more closely approximates nature's choice. An in depth exploration of this style has been the focus of Greene's theoretical work (Greene, in press, a; in press, b; Turvey, in press); in this paper we touch base with but a few aspects of the style, ideally, ones that convey the flavor of the major constraints.

In the style of coordination that is depicted by the second machine, the management of the kinematic degrees of freedom is brought within the realm of possibility by subdividing all the the biokinematic links participating in a movement into a small number of connected groups (muscle linkages) (cf. Gelfand, Gurfinkel, Tsetlin and Shik, 1971). We will refer to such linkages as coordinative structures (Turvey, in press) defined, generally, as a group of muscles, often spanning many joints, that is constrained to act as a unit.

In this style of coordination, acts arise from the interactions among coordinative structures. Grillner (1975, p. 296) expresses this aspect of the style: "...the simplest way to achieve a certain pattern from a group of different automata...would be to let them interact with each other, rather than to control each individual automaton from the outside...." To obtain different patterns from a set of coordinative structures one need only alter the mode of interaction. The modulation of coordinative structures and of the relationship among them is most commonly referred to as tuning (and for different kinds see Greene, 1972; Turvey, in press), and there are good reasons for assuming a devolution of responsibility for activating and tuning coordinative structures (Greene, 1972; Boylls, 1975; Turvey, in press). Insofar as tuning is linked to perception, we suggest the term coordinative process (in the interests of symmetry) and, maintaining the thrust of Bernstein's (1967) argument, claim for coordinative processes the decisive role in the achievement of motor control.

In the light of these last remarks, comes the conclusion to this elementary introduction to the degrees of freedom problem with rephrasing of our earlier question. How do coordinative structures and coordinative processes conflate to constrain the planning of an act and its evolution into a particular biokinematic trajectory?

#### Relating Postural and Transport Movements

Two classes of movement in gross motor tasks are distinguished: transport and postural (cf. Smith and Smith, 1962). Both classes may be regarded as transformations of posture, that is, configurations of trajectories as the limbs move from one relatively stable arrangement to the next, although transport transformations are often more intricate and sometimes more arbitrary than postural. The principal distinction between the two is that transport movements are orientations to the local conditions of stimulation, for example, the flight of the ball or the motions of an opponent, while postural movements are orientations to the global conditions of terrestrial stimulation, the global physical invariants (Shaw and McIntyre, 1974) such as the horizon, gravity, and the ground plane. It goes without saying that most gross motor acts--as manifest in tennis, soccer, and so on--involve a tight confluence between the two classes. Our question is: How is this confluence realized?

Fomin and Shtil'kind (1972) have introduced the term "pedate system" for any system with legs such that the system's normal contact with the surface of support is by means of the plantar parts of the feet. For nature's pedate



systems, surface contact through the feet, inertial contact through the vestibular system, and optical contact through the ocular apparatus are the three sources of information about the system's orientation and movement relative to the environment (Lee, in press). Of the three, vision is the more informative and influential; the vestibular system is not sensitive enough for fine balance control (Lee and Lishman, 1975), and surface contact through the feet is ambiguous about the body's relation to the environment when the feet move relative to the environment, such as when the surface is compliant, unsteady or narrow (Lee, in press).

From Gibson (1966) and others (Lee, 1974; Warren, 1976), we have learned that the optical flow patterns at the eye are specific to one's movements with respect to the layout of environmental surfaces. To illustrate, a person attempting to maintain an upright steady stance is perturbed by transformations of the total optic array: a form of inclusive optical expansion induces backward body sway, and a form of inclusive optical contraction induces forward body sway (Lee and Aaronson, 1974). As witness to the human pedate system's sensitivity to this visual source of exproprioceptive<sup>1</sup> information is the observation that body sway can be driven phasically by extremely small oscillations in optical expansion and contraction (Lee, in press).

We can claim, therefore, that while standing or locomoting, the maintenance of an upright posture is an active process (cf. Aggashyan, Gurfinkel' and Fomin, 1973) oriented principally (but not solely) to preserving the absence of certain kinds of inclusive optical change. Patently, any transport movement is, in the final analysis, a disturbance of the body's relation to the global invariants that are specified primarily by the optical flow pattern. For a great many transport movements, the movement is possible only if, during the movement, a relatively stable relation is preserved between the body as a unit and the global invariants. Could this be achieved simply by feedback, that is, by a process in which the perturbation is corrected for subsequent to the movement, or better still, subsequent to phases of the movement? The problem with a feedback solution is that the specified compensatory changes are often for states that are no longer current. The argument will be made that while some form of feedback (for example, velocity or acceleration feedback) is necessary to the integrity of the transport movement-postural movement relations, it is not sufficient. Consider, in this regard, the concept of "region of reversibility" as it relates to the concept of a pedate system (Fomin and Shtil'kind, 1972).

The set of all transformations of the biokinematic chains defines a phase space of which a subset is the region of controllable transformations. Within the latter, a particular subset is defined such that for any two points in the subset, there is a control process by which either point can be

---

<sup>1</sup>Lee (in press) suggests the term exproprioceptive for proprioceptive (in the classical sense) information about the orientation, position and movement of the body, or of a body part, relative to the environment.

attained from the other; in short, for any movement defined within the subset there is an inverse. This subset is the "region of reversibility," and by the use of the term "equilibrium" for a pedate system, we mean that the kinematic state of the system is within this region. Now it follows that a major constraint on the planning and executing of many transport movements is that they conserve the pedate system within this region. More precisely, and more practically, the constraint is that transport movements do not carry the system too close to the boundaries of the region. Proximity to the boundary is costly, in that coordinative effort would have to be disproportionately allocated to postural movements at the expense of transport movements.

The region of reversibility for a particular pedate system is not constant. Among possible sources of variation, we may recognize the conditions of the support surface and the speed at which the body is moving relative to the surface supporting its locomotion. Insofar as the region of reversibility for a given set of conditions is detectable, the question arises as to the nature of the information that specifies a region. Whatever the answer, one conjectures that for the region of reversibility concomitant to the conditions of a given skill, perceptual sensitivity to its boundaries is a determinant of the facility with which the skill is performed.

In summary, we have proposed that in addition to feedback, preserving a relatively invariant relation to the global invariants in the course of transport movements is partially achieved by an equilibrium-oriented constraint on the selection of transport movements. Consider a further possible factor.

As alluded to above, preserving balance through feedback alone would often be too late and too slow. This tardiness, however, can be circumvented. When a cat detects an incipient stumble, approximations to the proper muscular response are rapidly generated to preserve the upright posture of the cat long enough for relatively low-level feedback mechanisms to take charge (Roberts, 1967). A particularly sophisticated version of this style of control is suggested by the observations of Belenkii, Gurfinkel' and Pal'tsev (1967).

On receipt of an auditory signal, a participant is requested to raise his arm rapidly forward to the horizontal position. In the interval prior to the first signs of activity in the deltoid muscles of the shoulder, the muscles most responsible for the movement, there is evidence for considerable modification in the muscle states of the trunk and lower limbs. If it is the right arm that is raised, activity in the biceps femoris of the right leg and the sacrolumbar muscles of the left side precede activity in the deltoids. In addition, a definite anticipatory relaxation occurs in the left biceps femoris. We see, in short, an orderly pattern of change--of fixing and relaxing links in the kinematic chain of the body--preceding the transport movement of raising the arm. This pattern is both stable and specific to the transport movement: the pattern is constant over repetition and the pattern anticipatory to lowering the arm is distinctively different from that anticipatory to raising the arm (Belenkii et al., 1967; Pal'tsev and El'ner,

1967).

These anticipatory changes can be interpreted as intended to minimize the perturbations of the pedate system that would result from the movement of the arm. Insofar as these changes do occur prior to the movement, we may recognize the larger implication that, at least for this limiting case, the specification of a particular transformation of a kinematic chain that is a particular transport movement, is concurrently the specification of a particular transformation of other kinematic chains, which is the cognate, postural movement.

If these anticipatory postural adjustments are absent or impaired (owing to brain injury), then pronounced excursions in the center of gravity accompany the arm movement (Pal'tsev and El'ner, 1967). Nevertheless, the anticipatory adjustments are not the whole story, for in the normal case, other postural adjustments presumably of a more precise nature, accompany and follow the movement of the arm (Pal'tsev and El'ner, 1967). It seems as if the anticipatory adjustments put the pedate system into the ballpark (see Greene, 1972; in press b) of postural arrangements appropos the dynamics of moving the arm and appropos the disposition of the limb subsequent to the movement. To state the larger implication as noted above, more simply and somewhat differently: the plan for a transport movement, such as an arm motion, specifies the ballpark of necessary postural movements or, relatedly, a transport plan "pied-pipes"<sup>2</sup> an approximate, postural plan.

The preceding statement, in both its simple and more complicated forms, must be qualified on two counts. First, the relations between transport and postural movements is not as hierarchical as pied-piping would seem to imply. The weight of the evidence (Belenkii et al., 1967; Pal'tsev and El'ner, 1967; El'ner, 1973) favors a coalitional relation (see Turvey, in press). Second, there is the question of the generality of this form of control. There is the possibility, of course, that the balance-oriented fixing and relaxing of biokinematic links preparatory to and specific to a transport movement is manifest only in simple motor tasks such as studied by Belenkii, et al. (1967). We can argue on rational grounds, that the form of control described above would be apt for many forms of transport movements; as a general principle, approximating a desired state through feedforward makes the task of feedback regulation considerably more simple and more efficient (cf. Greene, 1972; in press a).

Suppose, therefore, that in the general case, the intended transport movement can be the basis for specifying an anticipatory but approximate feedforward adjustment of postural control structures. In the acquisition of a skill (say, a gymnastics routine), it would be beneficial for the performer to become sensitive to the postural movement implications of intended

---

<sup>2</sup>This term was suggested to us by Robert Remez, with all due respect to John Robert Ross.

transport movements (see Belenkii et al., 1967, for a modest demonstration). An advantage of this sensitivity is that by approximating postural controls ahead of time, the performer can devote more coordinative effort to the intricacies of the skill.

#### Coordinative Structures and Coordinative Processes

Whether our concern is with the transformations on the biokinematic chains that constitute transport movements, or whether it is with those that constitute postural movements, it remains the case that the movements are fashioned from the interplay of coordinative structures and coordinative processes. This section seeks to express something of the flavor of coordinative structures and the manner of their modulation. An initial step toward a more precise characterization is to look at some of the biokinematic events that have been (or may be) promoted as instances of the concept of coordinative structure.

The activity of a single limb during locomotion consists of two broadly defined phases: support and transfer. The support phase, during which the foot is in contact with the ground, is composed of extensor activity over the limb joints; the transfer phase, carrying the foot from one support to the next, is composed essentially of flexion. In an ingenious experiment (Orlovskii and Shik, 1965) conducted with dogs locomoting freely on a treadmill, a very brief impedance was applied at the elbow during transfer-flexion. Consequently, the movement at the elbow was slowed, but so was the movement at the shoulder and wrist. However, a similar impedance delivered during support-extension did not retard the movement at the other joints. It is arguable that the link motions during flexion are constrained to act as a unit by means of spinally-mediated afferentation (cf. Boylls, 1975). What of the extensors? They, apparently, are not linked by shared afference, though they do appear to be linked--that is, they do behave as a unit during locomotion. Witness to this claim is the observation that across various gaits, the timing of limb extensor EMGs is nearly invariant with respect to step cycle and, further, that the activity periods of extensor muscles relative to each other change little as speed of locomotion changes (Engberg and Lundberg, 1969). Perhaps the implication is that in locomotion the limb extensors are constrained to act as a unit by means of common efference (Boylls, 1975).

A unitary arrangement of joint changes that has been investigated quite thoroughly and which, therefore, provides an exemplary case, is that which preserves the stability of the head during respiration (Gurfinkel', Kots, Pal'tsev, and Fel'dman, 1971). With inspiration and expiration, the torso (in both its upper and middle parts) deflects backwards and forwards, respectively. The displacement is of sufficient magnitude so that, if left unchecked, marked excursion would occur in the overall center of gravity. However, the respiratory-induced oscillations in the torso are balanced by antiphasic oscillations at the hip and at the cervix. Changes in the angle of the hip and of the cervix are simultaneous with changes in the angle of the torso, and the relation among these changes is invariant with frequency



of respiration. This constraint on the biokinematic chain is wrought neither by means or mechanical conspiracy nor by spinally-mediated afferentation (Gel'fand et al., 1971); as with the extensors during locomotion, the coupling source is probably efferent.

The control of two joints of the arm illustrates a further case. When a person is requested to simultaneously flex or extend his wrist and flex or extend his elbow, the joints are moved mainly in a coupled fashion (Kots and Syrovegin, 1966), although his synchrony is achieved with less practice in the case of changes of the same type (for example, flexion-flexion), than in changes of the opposite type [for example, extension-flexion (Kots et al., 1971)]. It is significant that the two rates of change of joint angle preserve one or another invariant ratio that is not attributable to mechanical coupling. Individuals tend to have three to seven such ratios and to differ in the ratios they use. Furthermore, they use a different subset of these ratios (usually three or four of them) for each of the four combinations of flexion and extension (Kots and Syrovegin, 1966).

Finally, let us note observations on the production of speech which suggest that often movements of the tongue, lips, velum and jaw may be constrained as a unit (Kent, Carney and Severeid, 1974).<sup>3</sup> To illustrate, in uttering the word contract, lowering of the velum is initiated with the release of oral closure for /k/, and elevating the velum begins with the tongue tip movement for alveolar closure (Kent et al., 1974). In uttering the word we, the transition from the glide /w/ to the vowel /i/ is mediated by the contemporaneity of a forward gesture of the tongue body and a release of lip protrusion. With increase in emphatic stress, there is an increase in the displacement and velocity of the tongue body and in the displacement and velocity of the upper lip. However, the relation between the lingual and labial displacements and velocities remains invariant over variations in stress (Kent and Netsell, 1971). Apparently, for utterances like /wi/, the stress contrast modulates both articulators or neither articulator.

It is dimly apparent from these examples that where several muscles are synchronized as a unit, whether it be through spinally-mediated afferentation or central efference, the activities of the individual muscles covary in terms of a ratio that is indifferent to overall magnitude changes in these activities. In reaching this tentative conclusion, we are somewhat guilty of the "in-phase" assumption, for our examples have crossed the muscle state--link movement boundary--and we are treating the two as isomorphic. Nevertheless, we believe the conclusion has heuristic merit, and following Boylls (1975), we proceed to identify two "prescriptions" for a coordinative structure.

---

<sup>3</sup>We thank Carol Fowler for bringing these particular observations to our attention.



A "structural prescription" refers to the qualitative ratios of activities in the muscles composing a coordinative structure that are invariant with respect to absolute activity levels. A "metrical prescription" refers to the specification of the absolute activity level. As Boylls (1975) remarks, a metrical prescription is like a "scalar" quantity that multiplies by the same amount the activities of each muscle in a coordinative structure. In the above example, of the utterance /wi/, emphatic stress is analogous to a metrical prescription, for it magnifies the lingual and labial activities to the same degree; the ratio between the two activities that is preserved over stress is the structural prescription.

The specification of structural and metrical prescriptions for coordinative structures is, in part, what is meant by coordinative processes. A change in structural prescription changes the dynamic topography of a biokinematic chain whose links have been constrained to act as a unit; a change in the metrical prescription changes, among other things, the speed with which the dynamic topography is realized. We intuit that in the course of transport and postural movements, metrical prescriptions can be modulated more rapidly and with greater facility than structural; there are a few experiments in favor of this intuition.

From the work of Asatryan and Fel'dman (1965) and Fel'dman (1966), it is learned that where the muscles at a joint have been constrained to act as a unit--either for the preservation of a particular posture against opposing moments of force or for the purpose of moving, on signal, to a new prescribed position, again against opposing moments of force--the muscle complex can be described as a nonlinear spring with definite stiffness and damping parameters. In the case where a posture is to be maintained, if the opposing moments of force are unexpectedly changed, the limb segment moves initially to a posture that is in accord with the original parameters, and only then does it move to a posture that is in accord with the new parameters relevant to the new moments. In the case of moving to a prescribed position, if the moments are changed subsequent to the signal to move but prior to movement, the limb will move initially, but erroneously, to a position that would be predicted for the "spring" parameters present at the time of the signal. One might interpret these observations to mean that once a coordinative structure has been actuated, the parameters of that structure cannot be modified until the task, for which it was set, is complete. A more prudent interpretation is that the temporal scale over which changes wrought through coordinative processes can occur, does not always overlap the temporal scale over which changes wrought through generated kinetic energy can occur. Where the scales do overlap, the personality of a coordinative structure can change--in flight, as it were.

It can be shown by experiment (Vince and Welford, 1967) that a movement by a hand begun slowly in response to a signal for a "slow movement" can be accelerated in response to a further signal, one that is for a "fast movement," in very much the same time that it would take to initiate a fast movement from rest. This is so even if the second signal arrives during the latent period of the first. In this experiment and another (Megaw, 1970) in

which the second signal called for a slightly different movement from the first signal, it appears that the form of an "initiated" movement is less rapidly altered than the vigor with which it is conducted. In our terms, structural prescriptions are less rapidly alterable than metrical prescriptions.

An especially interesting illustration of metrical modulation is to be found in the activity of the baseball batter (Hubbard and Seng, 1954). In this illustration, the derivative properties of the optical flow field at the eyes provides the information for metrical prescription. As with all batting skills, it is mechanically advisable to move in the direction of the ball. The right-handed baseball batter does so by lifting his left and leading foot, moving it forward and parallel to the ball's line of flight, to finally place the foot some distance in front of, and probably slightly to the side of, the foot's initial position. The start of this step is synchronized with the release of the ball from the pitcher's hand. The duration of the step, however, and the start of the swing (which more often than not coincides with the completion of the step) are inversely related to the speed of the ball, to which the speed and duration of the swing remain relatively indifferent (Hubbard and Seng, 1954).

Consider the act of batting to be supported by a function defined over a small number of coordinative structures, that for present purposes suffices as our definition of an action plan (Turvey, in press). It can be hypothesized that the batter's stepping pattern arises primarily from the activities of knee extensors and hip abductors and flexors constrained to act as a unit. A structural prescription for this coordinative structure defines the dynamic topography of the stepping movement. The batting plan is initiated with release of the ball; as it unfolds, it is tailored to the current contingencies by the optically specified metrical prescription: the duration of the step (and hence the initiation of the swing) is functionally related to the speed of the ball.

This last example gives a glimpse of a central problem for the theory of how acting (coordinative structures) and perceiving (coordinative processes) conflate: in the performance of acts, exteroceptive, proprioceptive, and exproprioceptive information must be selectively percolated through the action structures at the right time. Conventional theories of selective attention do not address the question of how the selection of information is temporally constrained in order to be compatible with the dynamical requirements of the system it serves.

A quite remarkable observation of Orlovskii's (1972) may have some bearing on this problem. Given supraspinal stimulation of the spinal cord known to enhance flexor and extensor contraction in the inert animal, it was shown that when this stimulation was continuous with locomotion, the effects of the stimulation were manifest only at select points in the locomotory cycle. One might interpret this result as saying that the interaction of coordinative structures created "holes" or "slots" through which the continuously present supraspinal influences could "flow" (cf. Boylls, 1975). Is

this an instance of a general principle? In that the visual information that supports activity cannot be characterized as momentary signals or stimuli, but as continuous optical flow fields (Gibson, 1958; Lee, 1974), can it not be conjectured that the "introjection" of information into an act is constrained by the interaction among coordinative structures mediating the act? That selective percolation at the right time is defined, in very large part, by the act itself?

#### The Concept of a Control Surface for a Coordinative Structure

The problem of how perceptual information might be introjected into an act so as to appropriately constrain the manner in which it unfolds, requires some method for precisely defining the degrees of freedom of the biokinematic system that is involved in doing the act. We would especially like an answer to the question of how to determine the minimal number of degrees of freedom needed to control the coordinative structure, that is, to specify all of its configurations. For the present purposes, this question is raised only with respect to structural prescriptions. Our intuition is that an answer to this question may be systematically represented by a topological model of the kinematic chain referred to as a control surface (or space, as the case may be).

An illustration of a control surface can be provided by considering the guidance system required for a simple robot limb. Let the robot limb consist of two rigid segments of unequal length connected together by a joint (an "elbow") that permits  $360^\circ$  rotation in the plane, with this articulated limb connected by a similar circular joint to the main body of the robot (a "shoulder"). The minimal number of degrees of freedom needed to control such a system might be determined as follows;

The robot arm described above is, essentially, a compound planar pendulum with two degrees of freedom. In general, a precise representation of the total kinematic state-set of a mechanical system of  $k$  degrees of freedom can be provided by a control space defined over  $k$  parameters. The structure of the control space can be determined by taking the (topological) product over the unrestricted motions of the multiple linkages of its freely jointed kinematic chains. In the case of the idealized robot, the segment extending from the shoulder can rotate through a  $360^\circ$  planar angle around the shoulder joint, and similarly, for the lower segment connected from the elbow joint. Furthermore, since these two segments are independent, for each angle that one of them assumes, the other is free to assume any one of its continuous angular positions. Thus, all possible positions of this articulated limb may be represented as the topological product of angular positions of two circles. This topological product determines a manifold with two degrees of freedom known as a torus. The torus is a closed surface of two dimensions that resembles the outside of a bagel.

Intuitively, one can conceive of the topological product of two circles as that surface (a torus) generated by stringing a small circle on a large circle and then moving the small circle in such a way that the larger circle

consistently penetrates its center. All the points on the surface of this object specify possible kinematic states of the articulated limb of the robot. To locate a specific kinematic state on this surface requires imposing a coordinate system on the surface. Such a coordinate system is readily provided by dividing the circumference of the torus into degrees and, similarly, by dividing the circumference of the small circle into degrees. We can then use these two coordinate dimensions to locate every possible combination of circular joint values permitted by the kinematic linkages of the robot's limbs. Observe that the torus provides a natural model of the control space of the kinematic system of the robot because it represents every possible kinematic state and no impossible ones. A person might have thought that a surface like a sphere or an ellipsoid would do just as well, but neither of these is the product of variable-sized circles. The sphere can be made only with equal circles, and the ellipsoid is generated by a circle and an ellipse, not two circles. The torus is the only closed surface that can be generated as the topological product of two variable circles.

The above description is, of course, too idealized of a control space even for the robot. Neither of its joints will really allow a full  $360^\circ$  of rotation because the joints are restricted to the same plane. Thus, like the human arm, the robot's segmented limb will be restricted in its freedom of rotation. In general, the kinematic links of an animal or a human are constrained so as to restrict free variation--a fact that must be expressed in their control spaces. To illustrate, suppose that the shoulder joint of the robot permits free variation of its kinematic link through an angle limited to  $180^\circ$ , while the elbow joint permits free variation of its kinematic link through only  $90^\circ$ . Then the restricted control surface that represents these natural constraints is but one-quarter of the surface area of a half torus (cut in the way one would halve a bagel). In general, the natural constraint placed on the degrees of freedom of ideal kinematic systems can be represented as bounded portions of the ideal topological manifold corresponding to the control space of that system.

It is worth noting that the method of taking topological products over kinematic chains with more liberal joints will produce higher dimensional manifolds that represent control spaces of such systems. For instance, an articulated limb consisting of a ball joint and a  $360^\circ$  hinge-joint yields a topological product specifying a closed manifold of points--a control space--of three degrees of freedom, namely, the part of the space lying between two concentric spheres.

The theoretical reduction of the kinematic configurational states to a minimal representation as a control surface provides an ideal solution to the problem of determining the minimal degrees of freedom required for the control of a coordinative structure by a coordinative process. However, in so far as we may claim that the concept of a control surface is formally equivalent to that of a coordinative structure, the problem remains of how to represent the coordinative process required to control the coordinative structure. For the moment, only speculation as to the best way to formally represent this relationship can be made. Tentatively, we suggest that the



coordinative process interfaces with the coordinative structure by a moving point of control on the control surface whose control cycle over the surface provides a dynamic specification of both the order and rate at which the structural prescription for kinematic trajectories (acts) represented on the control surface are to be selected. But what controls the course and schedule of the moving point of control?

In the next section we attempt to provide an answer to this question by showing that the coordinative process that must guide the moving point of control has available to it the necessary information from the environment.

#### Conclusion: Coordinative Process and Coordinative Structure as Duals

We suggested above that the planning and executing of an act might be construed as the flow of a point of control over a control surface so as to selectively and progressively freeze out kinematic degrees of freedom. This flowing point of control, however, must be under the guidance of some other process--what we termed a coordinative process. In order to avoid a troublesome regress, we must determine where the coordinative process derives the constraints needed to accomplish this task. Briefly, our suggestion follows: the information that the coordinative process requires to constrain the degrees of freedom of the kinematic system left unconstrained by the coordinative structure is produced in the very process by which an act unfolds over time and space.

To illustrate this hypothesis, consider the following case: What kinematic capabilities must a robot possess to move across a room to a closed door and open it? First, the machine must orient to the door by rotating and then locomote in that direction by translating--a total of two degrees of freedom. Additionally, the robot's arm must possess at least one hinged joint and a prehensile organ attached to a rotating wrist joint so as to turn the door knob--an additional three degrees of freedom.

This kinematic system of five degrees of freedom must be controlled by some source with five degrees of constraint if the robot is to reach the door and open it. Where might the coordinative process find these degrees of constraint? To act felicitously, the robot must not only be endowed with action capabilities, but with perceptual capabilities as well.

Hence we would want the robot to be able to sense the following dimensions of its environment: the floor of the room, a plane with two degrees of freedom; a door with a knob at some determinate height, a third degree of freedom; a door knob of a certain size that rotates, a fourth and fifth degree of freedom respectively. Thus, an environment minimally complex for supporting the act of locomoting and door-opening must possess a specific set of values along five dimensions. Not coincidentally, each of these five values are precisely what is required to restrict the free variation of the robot's moving point of control on its control surface--a complex topological manifold of five degrees of freedom. Thus, as might be reasonably assumed, if the robot is capable of detecting the relevant dimensions of its



environment, then it follows that its coordinative process necessarily possesses a source of degrees of constraint which perfectly complement the degrees of freedom of its coordinative structure. Our conclusion follows immediately from the table below.

TABLE 1: Action/Perception Duals

Degrees of freedom (df) of the coordinative structure (CS) and the degrees of constraint (dc) of the coordinative process (CP) required for the act of crossing a room to open a door.

| CS                                               | <u>df of kinematic system</u>   | CP                                                   | <u>dc of environment</u>  |
|--------------------------------------------------|---------------------------------|------------------------------------------------------|---------------------------|
| $\beta_1$                                        | = 1 rotation of body            | $\phi_1$                                             | = 1<br>2 dimensions of    |
| $\beta_2$                                        | = 1 translation of body         | $\phi_2$                                             | = 1 ground plane          |
| $\beta_3$                                        | = 1 rotation around elbow joint | $\phi_3$                                             | = 1 height of door knob   |
| $\beta_4$                                        | = 1 closing of hand             | $\phi_4$                                             | = 1 size of door knob     |
| $\beta_5$                                        | = 1 rotating of hand            | $\phi_5$                                             | = 1 rotation of door knob |
|                                                  | <u>5</u>                        |                                                      | <u>5</u>                  |
| <u>df in action</u> = $\alpha: \beta_i + \phi_j$ |                                 | <u>dc from perception</u> = $\rho: \phi_j + \beta_i$ |                           |

Note: a felicitous action is simply an evaluation of the following linear relationships between the arguments of two functions,  $\alpha$  and  $\rho$ .

$$\begin{aligned} \alpha(\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5) &= \rho(\phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5) \\ \text{or} \quad (\beta_1 - \phi_1) + (\beta_2 - \phi_2) + (\beta_3 - \phi_3) + (\beta_4 - \phi_4) + (\beta_5 - \phi_5) &= 0 \\ (1-1) + (1-1) + (1-1) + (1-1) + (1-1) &= 0 \\ 0 &= 0 \end{aligned}$$

Hence degrees of freedom = degrees of constraint

In sum, we suggest that in all cases of felicitous action there exists a precise mathematical relationship between coordinative structures and coordinative processes, what might be called a duality under complementation such that the degrees of freedom of the former are perfectly balanced by the degrees of freedom of the other. In this way, the action/perception system provides sui generis its own control.

## REFERENCES

- Aggashyan, R. V., V. S. Gurfinkel', and S. V. Fomin. (1973) Correlation and spectral analysis of fluctuations of the human body during standing. Biophysics 18, 1173-1177.
- Aizerman, M. A. and F. A. Andreeva. (1968) Simple search mechanisms for control of skeletal muscles. Automation and Remote Control 29, 452-463.
- Arbib, M. A. (1972) The Metaphorical Brain: An Introduction to Cybernetics as Artificial Intelligence and Brain Theory. (New York: J. Wiley and Sons).
- Asatryan, D. G. and A. G. Fel'dman. (1965) Functional tuning of the nervous system with control of movement or maintenance of a steady posture - I. Mechanographic analysis of the work on the joint on execution of a postural task. Biophysics 10, 925-935.
- Belen'kii, V. Ye., V. S. Gurfinkel', and Ye. I. Pal'tsev. (1967) Elements of control of voluntary movements. Biophysics 12, 154-161.
- Bernstein, N. (1967) The Coordination and Regulation of Movements. (Oxford: Pergamon Press).
- Boylls, C. C. (1975) A theory of cerebellar function with applications to locomotion. II. The relation of anterior lobe climbing fiber function to locomotor behavior in the cat. COINS Technical Report 76-1, Department of Computer and Information Science. (Amherst: University of Massachusetts).
- Chernov, V. I. (1968) Control over single muscles or a pair of muscle antagonists under conditions of precision search. Automation and Remote Control 29, 1090-1101.
- El'ner, A. N. (1973) Possibilities of correcting the urgent voluntary movements and the associated postural activity of human muscles. Biophysics 18, 966-971.
- Engberg, I. and A. Lundberg. (1969) An electromyographic analysis of muscular activity in the hindlimb of the cat during unrestrained locomotion. Acta Physiol. Scand. 75, 614-630.
- Eshkol, N. and A. Wachman. (1958) Movement Notation. (London: Weidenfeld and Nicholson).
- Evarts, E. V., E. Bizzi, R. E. Burke, M. DeLong, and W. T. Thach. (1971) Central control of movement. Neurosciences Research Program Bulletin, 1971, no. 1.
- Fel'dman, A. G. (1966) Functional tuning of the nervous system with control of movement or maintenance of a steady posture - III. Mechanographic analysis of the execution by man of the simplest motor tasks. Biophysics 11, 766-775.
- Fomin, S. V. and T. I. Shtil'kind. (1972) The concept of equilibrium of systems having legs. Biophysics 17, 137-141.
- Gel'fand, I. M., V. S. Gurfinkel', M. L. Tsetlin, and M. L. Shik. (1971) Some problems in the analysis of movements. In Models of the Structural-Functional Organization of Certain Biological Systems, ed. by I. M. Gel'fand, V. S. Gurfinkel', S. V. Fomin, and M. L. Tsetlin. (Cambridge, Mass: MIT Press).
- Gellhorn, E. (1948) The influence of alterations in posture of the limb on cortically induced movements. Brain 71, 26-33.

- Gibson, J. J. (1958) Visually controlled locomotion and visual orientation in animals. Brit. J. Psychol. 44, 182-194.
- Gibson, J. J. (1966) The Senses Considered as Perceptual Systems. (Boston: Houghton Mifflin).
- Green, R. H. (1972) Problems of organization of motor systems. In Progress in Theoretical Biology, vol. 2, ed. by R. Rosen and F. M. Snell. (New York: Academic Press).
- Greene, P. H. (in press a) Strategies for heterarchical control--an essay. I. A style of controlling complex systems. International Journal of Man-Machine Studies.
- Greene, P. H. (in press b) Strategies for heterarchical control--an essay. II. Theoretical exploration of a style of control. International Journal of Man-Machine Studies.
- Grillner, S. (1975) Locomotion in vertebrates: Central mechanisms and reflex interaction. Physiological Review 55, 247-304.
- Gurfinkel', V. S., Ya. M. Kots, Ye. L. Pal'tsev, and A. G. Fel'dman. (1971) The compensation of respiratory disturbances of the erect posture of man as an example of the organization of interarticular interaction. In Models of the Structural-Functional Organization of Certain Biological Systems, ed. by J. M. Gel'fand, V. S. Gurfinkel', S. V. Fomin, and H. L. Tsetlin. (Cambridge, Mass.: MIT Press).
- Gurfinkel', V. S. and Ye. I. Pal'tsev. (1965) Effect of the state of the segmental apparatus of the spinal cord on the execution of a simple motor reaction. Biophysics 10, 944-951.
- Hubbard, A. W. (1960) Homokinetics: Muscular function in human movement. In Science and Medicine of Exercise and Sport, ed. by W. R. Johnson. (New York: Harper).
- Hubbard, A. W. and C. N. Seng. (1954) Visual movements of batters. Res. Quart. 25, 42-57.
- Kent, R. D., P. J. Carney, and L. R. Severeid. (1974) Velar movement and timing evaluation of a model for binary control. J. Speech and Hearing 17, 470-488.
- Kent, R. D. and R. Netsell. (1971) Effects of stress contrasts on certain articulatory parameters. Phonetica 24, 23-44.
- Kots, Ya. M., V. I. Krinskiy, V. L. Naydin, and M. L. Shik. (1971) The control of movements of the joints and kinesthetic afferentation. In Models of the Structural-Functional Organization of Certain Biological Systems, ed. by J. M. Gel'fand, V. S. Gurfinkel', S. V. Fomin, and M. L. Tsetlin. (Cambridge, Mass.: MIT Press).
- Kots, Ya. M. and A. V. Syrovegin. (1966) Fixed set of variants of interaction of the muscles of two joints used in the execution of simple voluntary movements. Biophysics 11, 1212-1219.
- Lee, D. N. (1974) Visual information during locomotion. In Perception: Essays in Honor of James J. Gibson, ed. by R. B. MacLeod and H. L. Pick. (Ithaca, N.Y.: Cornell University Press).
- Lee, D. M. (in press) On the functions of vision. In Modes of Perceiving, ed. by H. Pick and E. Saltzman. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Lee, D. N. and E. Aronson. (1974) Visual proprioceptive control of standing in human infants. Percept. Psychophys. 15, 529-532.

- Lee, D. N. and J. R. Lishman. (1975) Visual proprioceptive control of stance. J. Human Movement Studies 1, 87-95.
- Lieberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
- Lisin, V. V., S. I. Frankstein, and M. B. Rechtman. (1973) The influence of locomotion on flexor reflex on the hind limb in cat and man. Exp. Neurol. 38, 180-183.
- Litvintsev, A. I. (1968) Search activity of muscles in the presence of an artificial feedback loop enclosing several muscles simultaneously. Automation and Remote Control 29, 464-472.
- MacNeillage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev. 77, 182-196.
- McGaw, E. D. (1970) Response factors and the psychological refractory period. Unpublished thesis, University of Birmingham, England.
- Orlovskii, G. N. (1972) The effect of different descending systems on flexor and extensor activity during locomotion. Brain Res. 40, 359-371.
- Orlovskii, G. N. and M. L. Shik. (1965) Standard elements of cyclic movement. Biophysics 10, 935-944.
- Paillard, J. (1960) The patterning of skilled movements. In Handbook of Physiology: Neurophysiology, Vol. 3, ed. by J. Field, H. W. Magoun, and V. E. Hall. (Washington, D.C.: American Physiological Society).
- Pal'tsev, Ye. I., and A. M. El'ner. (1967) Preparatory and compensatory period during voluntary movement in patients with involvement of the brain in different localization. Biophysics 12, 161-168.
- Roberts, T. D. M. (1967) Neurophysiology of Postural Mechanisms. (New York: Plenum Press).
- Shaw, R. E. and M. McIntyre. (1974) Algoristic foundations to cognitive psychology. In Cognition and the Symbolic Processes, ed. by W. Weimar and D. Palermó. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Smith, K. U. and W. H. Smith. (1962) Perception and Motion. (Philadelphia: W. G. Saunders).
- Turvey, M. T. (in press) Preliminaries to a theory of action with reference to vision. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Vince, M. A. and A. T. Welford. (1967) Time taken to change the speed of a response. Nature 213, 532-533.
- Warren, R. (1976) The perception of egomotion. J. Exp. Psychol.; Human Percep. Perform. 2, 448-456.
- Wells, K. F. (1961) Kinesiology. (Philadelphia: Saunders).



## Physiological Aspects of Speech Production\*

Katherine S. Harrist†

### ABSTRACT

Speech production is a multilevel process, of which only the most peripheral stages are available to experimental observation. A central task of modern speech production research has been to study the interrelationships among these stages--the acoustic signal, the articulatory shapes generating it, and the muscle contractions controlling the movement of the articulators. A consequence of this work is that we may be able to infer the deviant aspects of abnormal production from their acoustic consequences.

### WHY STUDY SPEECH PRODUCTION?

Since the theme of this paper is the insight provided by modern research into language for the clinician working in the clinic or school, I should begin with a consideration of the characteristics of the clinician's patients. I will confine my attention to speech pathology, and hope that the problems faced by audiologists are adequately covered by others.

My own particular research interest is in the physiology of speech production, and so, presumably, the application of the research is to the speech production problems of the clinician's population. In recent years, a preoccupation with articulation and its disorders has become unfashionable among speech pathologists--interest has focused on the language- and

---

\*A version of this paper was presented at the conference "Implications of Basic Speech and Language Research for the School and Clinic," held at Belmont, Maryland, May, 1976, and will appear in Implications of Basic Research in Speech and Language for the School and Clinic (working title) to be published by MIT Press, Cambridge.

†Also The Graduate School and University Center, City University of New York.

Acknowledgement: This work was supported in part by a grant from the National Institute of Dental Research. I have benefitted from comments by Fredericka Bell-Berti, Gloria Borden, and Thomas Gay.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

learning-disabled child, and students are more excited by courses in psycholinguistics and syntax than in physiology and phonetics. While this trend probably cannot be reversed, I think we can make a case for the clinical importance of a continuing research commitment to these traditional topics.

In a few clinical populations, particularly the cleft palate group, the profoundly deaf, and the severely dysarthric, the patient's speech is so unintelligible that communication with the world is seriously impaired, and remediation of the symptom itself, by whatever means, is a problem of direct importance. While these populations are substantially large enough to arouse the attention of makers of health policy, most clients in the case load of the average speech pathologist are far less seriously handicapped. Stamping out voice disorders is not high on anyone's list of national health priorities. However, there is considerable evidence that the importance of voice and articulation disorders may be not only in the symptoms themselves, but also in their value as diagnostic signs for other disorders. Well-known examples are hoarseness as an early sign of laryngeal tumor, and slurred speech as a symptom of stroke. Learning more about the physiology and neurophysiology of speech production would enable us to make better use of speech symptoms as tools for the diagnosis of structural and neurological damage, as well as to improve symptom treatment per se.

As a field of study, speech production has a curious double nature. On the one hand, it is a complicated form of motor behavior, like walking, and can be studied by the same techniques of movement analysis. On the other hand, it is the output of the communication system, so that the movements themselves generate acoustic consequences, which must engage the perceptual capabilities of the listener in order to be effective.

#### LEVELS OF STUDY OF SPEECH PRODUCTION

Given that the study of speech production is a topic of general interest, it is perhaps worthwhile to consider how speech might be studied by considering an elementary description of the speech production process itself.

The listener, in hearing speech, operates upon an acoustic signal that has been generated by an extremely complicated system. Figure 1 shows a schematic diagram of this system.

Consider the system to be comprised of the three parts shown in Figure 1. The system is driven by energy supplied by the lungs during expiration. When inspiration takes place, the lungs, a pair of elastic, balloonlike structures, are filled with air. Since the walls of the lungs are elastic, they will tend to deflate to a neutral condition, and air can, in addition, be forced from them by the action of those muscles that control expiration. The larynx acts as a valve that can be moved into place over the expiratory air stream to convert the relatively steady expiratory air flow into puffs of air. The sequence of puffs of air, the volume velocity waveform, excites the

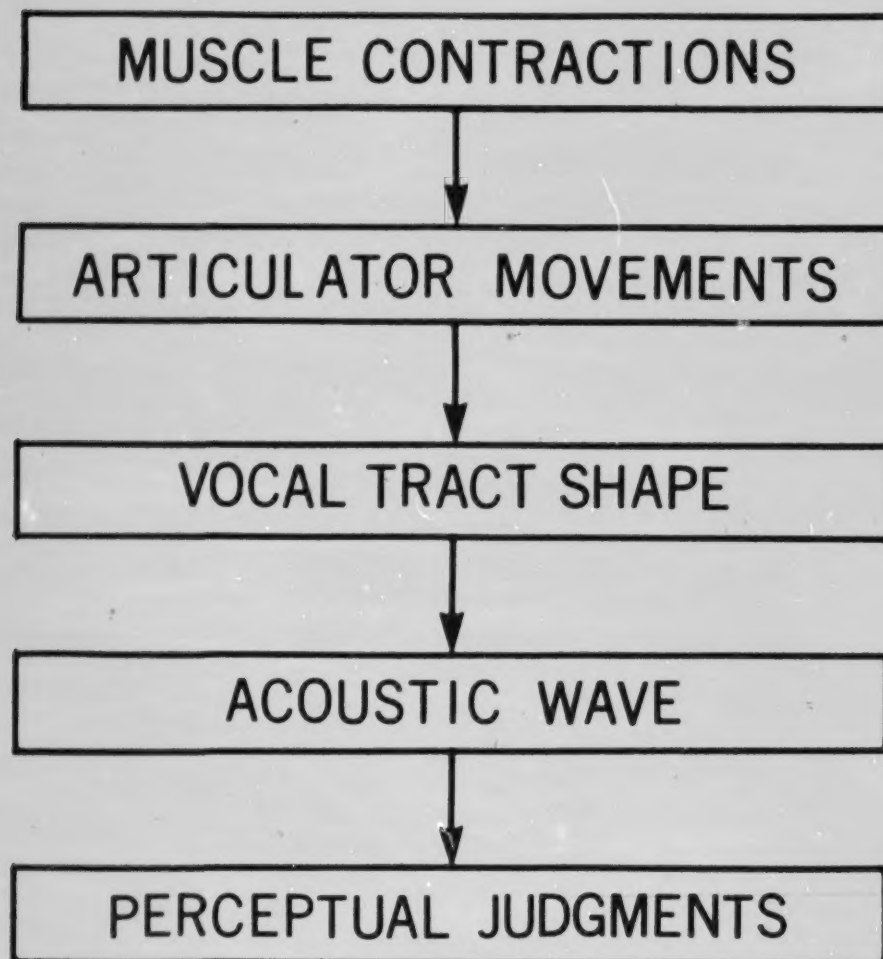


Figure 1: Schematic diagram of the articulatory system.

upper vocal tract, the oral and nasal air spaces, the single or branched tubes lying above the larynx. When the velum is lowered, air flows out the nose and mouth; when it is elevated, the nasal air space is closed off so that air flows out the mouth only. The upper vocal tract acts as a variable acoustic filter on the output of the larynx; its properties as a filter depend on its shape. The shape is controlled by the movements of the articulators. The velum is a muscle mass that acts to connect or shut off the nasal branches of the upper vocal tract from the air space below. The nasal air space has an almost constant shape. On the other hand, the oral air space has a shape that varies considerably depending on the position of the various articulators, such as the tongue, jaw, and lips. The position of the articulators, in turn, is controlled chiefly by the complex interactions of a large number of muscles.

This whole system can be studied at a number of levels, as indicated in Figure 2 (which can be considered an enlargement of the lower segment of Cutting and Pisoni's Figure 1). While these levels are closely related to each other, the relations between them are complex. In order to simplify the material to be discussed, I will not discuss aerodynamic factors or respiratory phenomena.

Speech disorders are probably most often studied by perceptual means--that is, the observer, or therapist, listens to the speech of the subject, or client, and makes some kind of judgment about the speech. The most usual kind of judgment is about the phonetic string that the speaker has conveyed; at a less analytic level, the listener could make a gross judgment that the speech was normal or defective, either overall, or on some kind of unit-by-unit basis. We will comment on the problem of making judgments of abnormal production by perceptual means later on in this paper. With respect to making judgments about normal speech, we are discussing the characteristics of speech perception generally (as is discussed in Cutting and Pisoni's paper elsewhere in this report).

Referring again to Figure 2, the next level at which speech can be studied is the level of the acoustic signal. The acoustic signal conveys a great deal of information about the speech production process itself through analytic techniques that have been developed quite recently, and are best summarized in Fant's monograph (1960). Briefly, the acoustic signal can be considered to be the product of a source function and a filter function. The source function can be directly related to the behavior of the vocal folds, while the filter function can be related to the shape of the upper vocal tract.

A number of authors have suggested the use of the acoustic analysis for the detection of laryngeal pathology. The technique is to use analysis procedures to reveal the source function uncontaminated by the acoustic filtering effects of the upper vocal tract. The source function can then be used to infer the behavior of the vocal folds. Variants of this procedure have been proposed by Lieberman (1963) and Rothenberg (1973), by Koike and Markel (1975), and by Davis (1976) as a technique for early detection of laryngeal pathology; large scale clinical trials of these techniques still lie in the future.



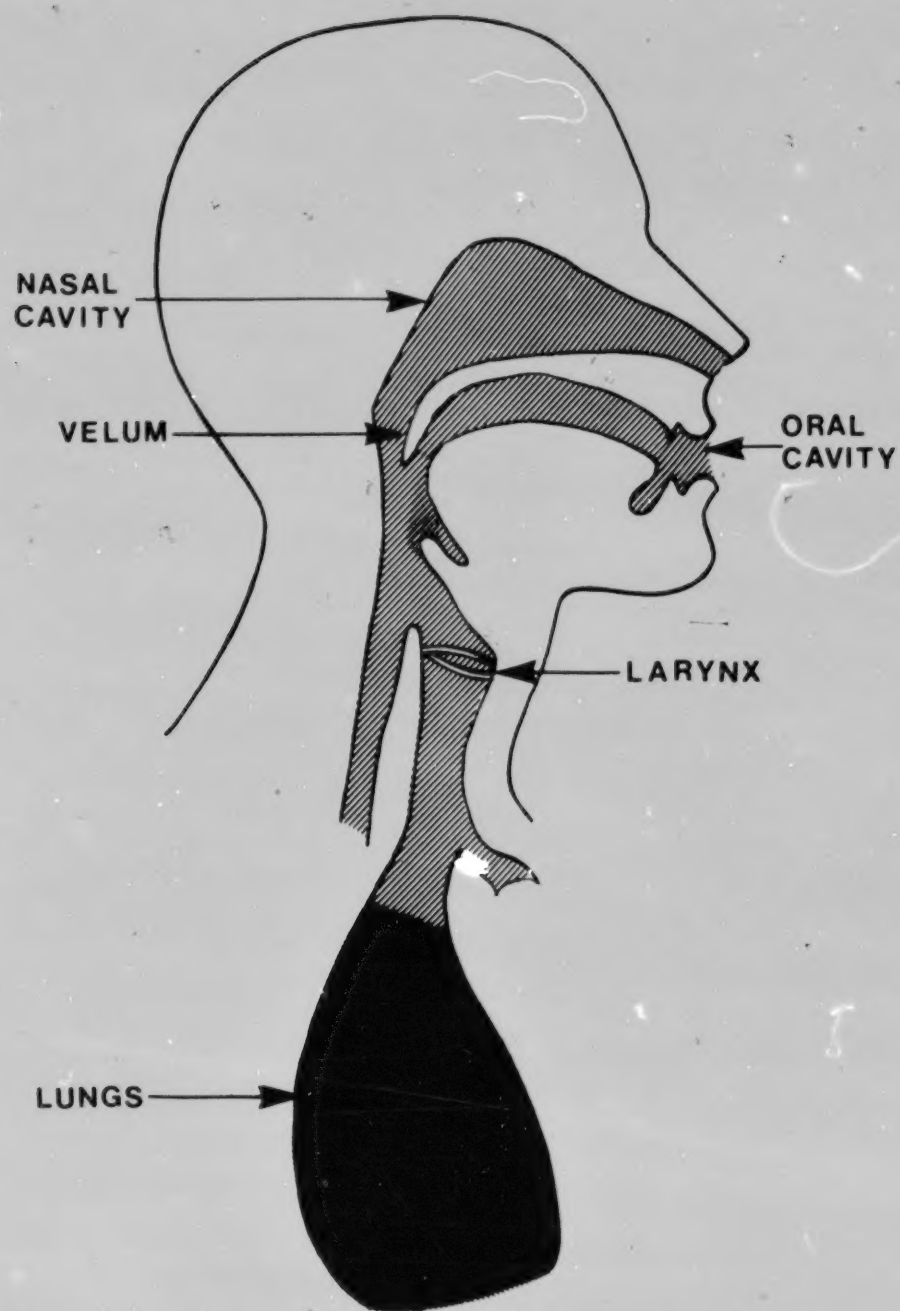


Figure 2: The last stages of the production side of the speech chain.

Of course, simpler types of acoustic analysis of abnormal voice production have been in widespread use for some time, by the analysis of such factors as the average pitch of patients or clients of various types [for example, Canter's study of the speech characteristics of patients with Parkinson's disease (Canter, 1963)].

The filter function, or the positions of the resonant frequencies of the vocal tract, are fairly well represented in a spectrographic display, although additional analysis procedures, such as linear predictive coding (Markel and Gray, 1976) can be used to estimate the resonant frequencies directly. The relationship between vocal tract shape and resonant frequencies has been the subject of study since the time of Helmholtz, although progress has been far more rapid since the Second World War [see Fant (1960) and the references therein]. There are two major limitations on the use of acoustic techniques for estimating vocal tract shape. First, in spite of great strides in relating shape to filter function, there are still some unknown factors preventing the verification of a complete model--such as acoustic losses due to absorption in the tract walls. A more serious problem is that the relationship between shape and acoustic output is inherently assymetric--the shape determines the output, but a given output may result from more than one shape (Stevens and House, 1955). Furthermore, a given shape may result from several equivalent articulatory postures (Lindblom and Sundberg, 1971). Therefore, even if the shape could be related unequivocally to the formant positions, an acoustic analysis is an inherently incomplete substitute for physiological observation.

Acoustic analysis has not been widely used as a technique for studying abnormal articulatory movement. There are two general and rather preliminary studies of the dysarthrias using spectrographic analysis (Lehiste, 1965; Lindblom, Lubker, and Fritzell, 1974).

These studies show the usefulness of spectrographic analysis for inferring such articulatory faults as restricted range of articulatory movement, as shown by vowel neutralization; abnormal rate of overall articulator movement; and abnormal control of the timing of the onset and offset of voicing. The same types of observations have been made, from spectrographic analysis, for the speech of deaf talkers. Angelocci, Kopp, and Holbrook (1964) have shown that the vowels of deaf talkers are severely neutralized. The timing of articulatory movement is grossly abnormal (Monsen, 1974), as is transitional movement from consonant to vowel (Monsen, 1976). In spite of the limitations of the inferences that can be made, acoustic techniques probably deserve more exploration than they have received, particularly because of their totally noninvasive nature.

Referring again to Figure 2, the movement of the articulators themselves can be studied. This is difficult, because except for the lips and jaw, they are extraordinarily difficult to observe directly.

Direct viewing of the vocal folds as a technique for the study of laryngeal pathology, was, of course, pioneered by the study of high-speed movies of Moore (1938) and von Leden, Moore, and Timke (1960). While this work has been useful, the use of the laryngeal mirror makes it impossible to

study anything other than sustained phonation, and the task of analyzing any significant number of cases by this technique is laborious due to the necessity of measuring the fold movement on a frame-by-frame basis. Recently, attempts have been made to automate the latter procedure.

The limitations on the use of the laryngoscope have been partly overcome by the development of the flexible fiber laryngoscope (Sawashima and Hirose, 1968). While it allows the viewing of the larynx during running speech, it does not transmit enough light for examination of single cycles of laryngeal vibration. However, good use of the technique has been made by McCall, Skolnick, and Brewer in videofluoroscopic examination of the larynx in spasmodic dysphonia (McCall, Skolnick, and Brewer, 1971).

The flexible fiberscope can also be used to view the velopharyngeal port, and has been so used for studying velar height during closure by Bell-Berti and Hirose (1975) and by Ushijima and Hirose (1974). Similar techniques probably will be useful as a substitute for x-ray movies now used for assessing the adequacy of cleft palate repair.

Probably the most important articulator is the tongue, and it is also the most difficult to view directly. The only practical, widely used method for viewing the shape of the upper vocal tract and the movement of the articulators is cinefluorography.

In the technique, x-ray motion pictures are taken of a sagittal view of the head, usually with salient points marked with pellets. In later analysis, the movies are projected frame-by-frame, and the movement of the articulatory structures tracked by the marking of the pellet positions in successive frames (Houde, 1967; Kent, 1972). There are both obvious and not-so-obvious limitations of this technique. First, the analysis is extremely time-consuming; although computer techniques can be used to simplify the measurement problems. Second, a given subject can only be used for a few minutes every six months, because of the dangers of exposure to excessive radiation. This limitation precludes the use of a number of niceties of experimental design, such as repeated use of the same subject, so that there has been inadequate exploration of differences both from individual to individual and from condition to condition.

In spite of these limitations, the description of normal or abnormal articulatory movement taken from x-ray movies is probably the most widely used technique in speech production research today. Much of the development of the technique can be ascribed to the efforts of Moll and his coworkers. In clinical research, cinefluorography has long been a standard way of assessing the adequacy of velopharyngeal function after cleft repair [see for example, Moll (1960); Subtelny, Koepp-Baker and Subtelny (1961)]. More recently, the same methods have been used in studying the dysarthrias (Kent and Netsell, 1975; Kent, Netsell, and Bauer, 1975; Netsell and Kent, 1976).

Again referring to Figure 2, we see that one can study not only the movement of the articulators, but even the control signals that lie behind the movements, by studying the electromyographic (EMG) signals to the muscles.

In speech research, two slightly different approaches have been taken to the recording of EMG signals. The difference between the approaches can be better understood by discussing the relevant EMG signal at slightly greater length.

The chief cause of articulator movements is the contraction of the attached muscles. Each muscle is made up of a number of fibers that contract by longitudinal sliding of their constituent myofibrils. When this shortening occurs, electrical changes that can be picked up by appropriate electrodes and amplifiers, occur in the muscle. Muscle fibers are grouped into motor units, each consisting of a neuron with its axon and the muscle fibers that it supplies. The muscle fibers of a given motor unit contract almost simultaneously, and the recorded electrical changes will consist of a complicated series of positive and negative waves whose precise size and shape will depend on the physical relationship of the recording electrode and the firing fibers. The fibers of adjacent motor units overlap spatially, so that any electrode may record from one or several motor units. As the force of contraction increases, the interpulse intervals of a given motor unit action-potential train decrease (DeLuca and Forrest, 1973); and other motor units are recruited. Thus, as the force of contraction increases, the spikes visible in a recording from an electrode typically become both larger and denser. The recording from several motor units is called an "interference pattern." In order to deal with this pattern quantitatively, it is customary to rectify and integrate it. An early study shows that the amplitude of integrated EMG is linearly related to muscle force in isometric contraction (Bigland and Lippold, 1954), and the relationships between force and EMG measures under various other conditions has been a subject of continuing study (Bouisset, 1973). For the speech musculature, the relationship between muscles and movement is so complicated that an analytic quantitative model for the relationship seems inherently unlikely. However, it has been shown that there is a monotonic relationship between EMG amplitude of the levator palatini muscle and palate height (Bell-Berti and Hirose, 1975) and between amplitude of activity in the posterior crico-arytenoid muscle activity (the PCA is an abductor of the vocal folds) and the size of laryngeal opening (Hirose, 1975), both during selected speech tasks. Thus we might assume the relationship between EMG amplitude and articulator movement for some sort of increasing function, although the detailed shape of the relationship may be experimentally impossible to establish.

The two types of approach to EMG signals are different because the goals are different. On the one hand, one may be interested in the nervous control of individual muscle units in the speech musculature, in normal or abnormal conditions. For example, MacNeilage and his associates (MacNeilage, 1973; MacNeilage, Sussman, and Powers, in press) have been interested in characterizing the various muscles of the orofacial area with respect to their action-potential trains; for this purpose, single motor units must be recorded, with an electrode with minimum field size. The corresponding clinical work is reflected in any standard text on clinical electromyography. There have been few studies of this general type with respect to the speech musculature, except for the few studies of lip muscles (Leanderson, Persson, and Ohman, 1970; Netsell and Cleeland, 1973; Netsell, Daniel, and Celestia, 1975). In general, the technique has been used to show generalized



hypertonia or muscle weakness.

One may also be interested in electromyographic signals from a kinesiographical point of view, that is, one may be interested in the forces acting on an articulator to cause a given movement for normal or abnormal cases; for work of this sort, one wishes to use a recording technique that will characterize the behavior of the whole muscle.

As indicated above, the relationship between EMG signals in any form and articulator movement is complicated. What is less commonly recognized, is that it is as difficult to specify important muscles for a given movement from an observation of the movement, as it is to move in the other direction. In part, this difficulty is due to the lack of anatomical studies of the orofacial region. In part, it is due to the fact that what appears on anatomical grounds to be a possible mechanism for muscular control of a structure, may not be so used. Furthermore, the speech structures are tied together in such a complex way that it is often impossible to distinguish active from passive movement by observation of structural movement alone.

Questions about the mode of closure of the velopharyngeal port to shut off the nasal air space provide a good example of this point. The velopharyngeal valve is surrounded by muscles, whose primary purpose is to block the opening between the oropharynx and the nasopharynx.

The levator palatini, the major muscle of the velum, can close the port by moving the velum up and back. The muscle of the upper pharynx, the superior constrictor, could close the port by pulling the posterior pharyngeal wall forward; the two muscles of the faucal pillars, the palatoglossus and the palatopharyngeus, could contract to pull the velum down, or could, on anatomical grounds, act with the other muscles to close off the port in a sphincter fashion.

EMG studies of the velopharyngeal mechanism have been undertaken by several investigators to determine, for speech articulation, which muscles are active in closing the port and which muscles are active in its opening. Fritzell (1969) concluded that the levator palatini is the most important muscle involved in closure in normal speech, and that superior constrictor activity was similar to that of the levator palatini; palatopharyngeus activity was different from individual to individual; and palatoglossus was active for nasal articulations, lowering the velum. Lubker, Fritzell, and Lindqvist (1970) confirmed Fritzell's (1969) original report that levator palatini activity, for speech, is highly correlated with velopharyngeal closure, and that palatoglossus activity is highly correlated with velopharyngeal port opening. Bell-Berti (1976) has also confirmed Fritzell's finding that the levator palatini is the most important muscle involved in velopharyngeal closure in normal speech. However, Bell-Berti's results do not confirm to Fritzell's on the superior constrictor and palatopharyngeus, and Lubker et al.'s results on the palatoglossus. She reported no consistent pattern of EMG activity in the superior pharyngeal constrictor, but found palatopharyngeus activity to be very similar in pattern to levator palatini activity. Palatoglossus activity was related to tongue body and lateral pharyngeal wall movement, rather than to velar lowering; velar lowering is

thus assumed to result largely from relaxation of the muscles of velopharyngeal closure.

There is no evidence in Bell-Berti's results that the superior constrictor contributes in a substantial way to velopharyngeal closure in a sphincteric fashion, as suggested by Skolnick, McCall, and Barnes (1973). However, she examined only three subjects, and a larger sample may reveal significant individual differences in this regard. The point here is that a possible mechanism is not a necessary mechanism.

The above summary is intended to show the interrelationships and overlap between various possible levels of study of the articulatory system. Ideally, any articulatory disorder should be discernible at every level, but a number of practical considerations stand in the way. In addition to the problems referred to above, most physiological techniques require a cooperative, or even a brave subject, which diminishes the usefulness of some patients, and of small children. Acoustic techniques are totally noninvasive, but may require equipment not available in the clinic setting. Therefore, perceptual techniques have been widely used, in spite of their incompleteness. Before I return to the perceptual techniques generally, however, I would like to discuss the argument that one of the other levels in the production process has some kind of primacy, in terms of the goals of the production process. The argument is usually made with respect to either lack of contextual variability of the units at some level, or retrainability. After considering these arguments, we will move on to describe the perception techniques themselves, and how they might be made more useful.

#### THE GOALS OF ARTICULATORY ORGANIZATION

Another reference to Figure 2 makes it clear that the primacy argument has arisen out of views of the nature of the speech perception process. These views have become noticeably vaguer as time has passed. A comparison of two quotations will make this point clear. The first, (Liberman, Cooper, Harris, and MacNeilage, 1962) states: "Given the complex, highly encoded relation between phonemes and sound, and the more nearly one-to-one correspondence between phoneme and articulation, we have been led to assume that a reference articulation might well be a stage in the perceptual process...Consideration of how the speech production system works...makes it clear why the acoustic signal should be rather complex and remote from the phoneme, and, further, why the simplest relation to the phoneme might be found in the motor commands that actuate the articulators." A recent formulation is given by Cutting and Pisoni, who quote Lashley in saying: "The processes (of perception and production) have too much in common to depend on wholly different mechanisms." While it is oversimplifying both points of view to leave the quotations to stand alone, the notable thing about the comparison is that while the view of the relation between perception and production remains quite similar in the two quotations, the statement as to the locus of the interaction vanishes in the second; the further history of the hypothesis with respect to speech perception is summarized elsewhere. We will consider only what might be called a motor theory of speech production.

The production component of the motor theory was a Hypothesis that one level of the speech production process, the motor inputs to the muscles involved in generating a phonetic segment, are more nearly invariant than is either the articulatory or the acoustic target. The basis for the assumption was the spatial and temporal smear resulting from the recombination of adjacent elements. While some early EMG studies showed invariance of signal in varying environments (for example, MacNeilage, 1963; Harris, Lysaught, and Schvey, 1965), the notion of motor invariance at the muscle contraction level was strongly challenged by the combined EMG and cinefluorographic study of 36 CVC monosyllables by MacNeilage and DeClerk (1969).

In many ways, their conclusions have been more influential than their observations "In every possible case, some aspect of the motor control (that is, electromyographic signal) of a later syllable component was influenced by the identity of a previous one. Except in a few cases, some aspect of the motor control of an earlier syllable component was influenced by the identity of the following one." This work led MacNeilage in a later paper (MacNeilage, 1970) to conclude: "Paradoxically, the main result of the attempt to demonstrate invariance at the electromyographic level has not been to find such invariance but to demonstrate the ubiquity of variability....the essence of the speech production process is not an inefficient response to invariant central signals, but an elegantly controlled variability of responses to the demand for a relatively constant end." At this level, MacNeilage's formulation, as he recognizes himself, is rather like Tolman's criticism of the S-R theorists' account of the rat running a maze. The rat has learned the location of the goal box, rather than a chain of turning responses; he will reach the goal box by swimming, a wholly new set of responses, if the maze is flooded.

MacNeilage and DeClerk's work was done with surface electrodes (the technology of the period), so that it is difficult to know what specific aspect of the muscle signals were being examined. However, a careful examination of the EMG data shows that while some signals were influenced by the adjacent phone, some were not. Thus, while the paper counters the argument that a motor command representation contains much less variability than an articulatory position representation, their data and that of others show variability with context at both levels, and does not give a clear picture of when context effects may be expected, and to what degree. Although the paper suggests an articulatory target formulation, it does not specify the rules governing failure to achieve that target.

An alternative to either invariant muscle control signals or invariant articulatory targets, is that the speaker tries to reach acoustic targets; this point of view has been suggested by Ladefoged (1967, 1972) and by Lieberman (1973) and Stevens (1973). Ladefoged's views are based on two studies of vowel production. In the first study, he showed that speakers who believe they were producing equivalent articulations, in the form of equivalent cardinal vowels, were producing equivalent acoustic vowels. In the second study, he showed that different speakers, when asked to produce the "same" vowel, will use widely different combinations of tongue and jaw displacement, thus getting the same tract shape for different articulatory maneuvers.

Lieberman (1973) and Stevens (1973) have pointed out that at some points in the vocal tract, a small change in the position of the constriction will have a large effect on the acoustic output, while at other points, it will have relatively little. The positions corresponding to the "point" vowels, /i/, /a/, and /u/, are relatively stable. Thus, their acoustic signals are likely to be stable even if the articulatory target is not quite reached, either due to sloppy articulation or coarticulation. Furthermore, at a dynamic level, articulatory movement for a vowel might enter a quantal region sometime before--and remain there until--sometime after--the articulatory target was reached; although the articulators might be moving steadily, the acoustic vowel would thus be relatively stable.

While these points of view have been much discussed, there remains very little solid information, other than that already mentioned, about the relative amounts of coarticulation at acoustic, movement, and control signal levels. It seems worthwhile to review a few experimental studies together with a review of the implications of the studies for retraining.

#### Coarticulation

At an acoustic level, Lindblom (1963), Stevens and House (1963), Guman (1965), Stevens, House, and Paul (1966), Gay (1973), Bell-Berti and Harris (1975), and Ohde and Sharff (1975) have all reported coarticulatory interactions in speech acoustic signals. For the latter two studies, it is possible to compare the magnitude of anticipatory and carryover effects of consonants on an intermediate vowel; in general, carryover effects are more substantial.

At an articulatory level, interest has been concentrated on showing the number of segments over which coarticulatory effects spread, and the effects of such markers as word boundaries on coarticulatory spread. The most often quoted results show that anticipatory (Daniloff and Moll, 1968), (Benguerel and Cowan, 1974) and carryover (Dixit and MacNeilage, 1972) coarticulation may spread over as many as four segments, although the limitation of coarticulatory effects have not been carefully specified. This work is summarized, and the theoretical implications discussed, by Daniloff and Hammarberg (1973). One point has received perhaps less attention than it should have. As was pointed out above, the same acoustic result could, in some instances, be achieved by different articulations, yet there is little evidence that different articulators are used for a given acoustic effect when context changes. For example, a given degree of mouth opening might be generated by tongue or jaw, depending on context. There is no evidence I know of that this happens.

At the level of muscle contractions, context effects, particularly carryover effects, are very substantial as MacNeilage and DeClerk noted. Part of the reason for this is the nature of the muscle signals that occur in speech. In general, the magnitude of EMG signals is related to articulator movement. Therefore, the signal will be related to the distance between articulator positions; for example, the activity of the muscles which raise the jaw will be greater for a given consonant after an open vowel than after



a closed vowel. If the EMG signal for a given phone is different in two phonetic environments, then the position of the associated articulator was probably different for the proceeding phone. If the EMG signal is the same, the position was probably the same (if rate of articulator movement did not change). Again, analysis of the muscle signals could be used to establish that a different path is used to achieve the same vocal tract shape in different environments; there is no very convincing evidence of this type of reorganization. Summing up, there seems to be no reason to choose one level of articulatory organizations over another as showing less coarticulation. It can be shown to occur at all levels to which we have experimental access. There may be theoretical value to posit invariant signals at some level prior to the motor units, but such speculations are not open to experimental check.

### Retraining

The specification of the goals of articulatory organization has both theoretical and practical significance. If the speaker stores some kind of internalized spatial or acoustic target as the representation of a phonetic unit, then interference with the articulation should be followed by rapid articulatory adjustment, whether the change is in the sensory feedback that acts to guide articulator placement, or in the articulatory structures themselves.

With respect to the importance of acoustic feedback for retraining, the evidence is ambiguous. It is well known that children who are profoundly deaf are unlikely to develop normal speech, but the effects of later adventitious deafening do not seem to be as severe. People who have developed speech normally will continue to speak intelligibly, with rather modest effects on pitch and intensity, under high noise level conditions (Ringel and Steer, 1963), although the noise levels that can safely be used may not mask all feedback. Delayed auditory feedback has well known devastating effects on articulation, although the nature of the articulatory failure is not well understood.

A large number of studies have aimed at showing the effects of interfering with tactile and kinesthetic feedback from the articulators. The largest group consists of the nerve block studies of Ringel and his associates (Ringel and Steer, 1963; Scott and Ringel, 1971a, 1971b; Putnam and Ringel, 1972; Horii, House, Li, and Ringel, 1973). In these studies, dental anesthesia blocks were administered which were considered to interrupt the sensory feedback from the articulators, and hence interfere with correct articulation. More recently, it has been shown that the nerve blocks that were used are likely to have interfered with motor, as well as sensory, innervation of the articulators (Borden, Harris, and Catena, 1973; Abbs, Folkins, and Sivarajan, 1976); hence, the results using this technique are difficult to interpret, and further research along these lines seems unprofitable, unless a different technique can be developed.

A more promising approach to both theoretical and applied problems is the direct investigation of the effects of articulatory interference. In an

applied sense, these experiments provide evidence as to speech adaptation after dental or surgical reconstruction. In spite of the obvious practical nature of this work, there is very little objective or quantitative information on speech adaptation (Hamlet, 1973). Some investigators have offered essentially anecdotal observations of reorganization as proof of target theories of production. For example, MacNeilage (1970) has pointed out that pipe smokers can produce essentially normal articulations with a pipe clenched between the teeth, eliminating normal jaw movement. Nooteboom (1970) points out that many speakers can produce the acoustic equivalent of the vowel /u/, which is normally produced by rounding the lips, by lowering the larynx without lip rounding. The vocal tract is thus changed in the appropriate way by adding length to the larynx end instead of the lip end. He considers this as evidence for an acoustic, rather than an articulatory, target. Experiments of a more formal kind have been of two types. In one type, an intermittent interference is made to an articulatory movement. Smith and Lee (1972) and Folkins and Abbs (1975) have conducted preliminary experiments of this sort. In the latter experiment, loads were applied abruptly and without warning to the jaw during closure for the bilabial stops; speakers had no difficulty in making appropriate adjustments of lip activity so that appropriate closure was made. In the second type of experiment, a prosthesis of some type is put in the mouth, and the course of adjustment is studied. For example, Amerman and Daniloff (1971) used listener judgments to evaluate speech that was produced while the speaker was wearing a prosthesis; by this criterion, the vowels were judged to be normalized within five minutes of insertion; however, careful acoustic measurements were not made.

A more thorough set of studies on speech adaptation has been made by Hamlet and Stone (1976). They collected phonetic, acoustic, and physiological measurements from subjects wearing alveolar-palatal dental prostheses. The nature of the compensations resulting from the prosthesis varied considerably between subjects. Compensatory patterns were evident within 15 minutes of insertion, but defective speech was observed immediately after insertion of the prosthesis, after wearing it for a week, and 15 minutes after removal. Consonant defects were easily observed; acoustic analysis revealed vowel formant shifts of which the subjects were unaware.

Summarizing the altered articulation experiments, it becomes clear that opinions differ substantially as to the speed with which a speaker is able to adjust to a structural change. One has the impression that the more carefully the motor behavior is studied, the longer are the estimates of adjustment time. There is not at present an experiment that provides unequivocal evidence of acoustic or articulatory loci as the goal of articulatory organization, by a convincing demonstration of immediate adjustment to an alteration.

#### SPEECH PERCEPTION AS A TECHNIQUE FOR THE STUDY OF SPEECH PRODUCTION

Returning to perceptual techniques as the way of looking at speech perception that the clinician ordinarily uses, for practical reasons I would

like to discuss some types of systems now in use, and what kind of research might make them more useful. Of course, one can argue that perceptual adequacy is the ultimate test of normal speech. However, clinicians sometimes feel that perceptual techniques are useful beyond this level, in classification of deviant speech, or in therapy. Two rather different kinds of techniques probably deserve some special comments.

The first is the use of descriptors. A well-developed use of this technique is that of Darley and his associates (Darley, Aronson, and Brown, 1969a; 1969b; 1975). In these studies, the speech of groups of patients was ranked on a seven-point scale with respect to a large number of descriptors; the rankings for all patients were later factor analyzed. The disorders were then described by the highest ranking (most deviant) descriptors. For example, the four most deviant dimensions for patients with cerebellar lesions are "imprecise consonants, excess, and equal stress, irregular articulatory breakdown and distorted vowels." If the descriptors are easy to apply, statistically reliable, and discriminate well among syndromes, they then are clinically useful for the classification of new cases.

A limitation of the technique is that there is no tight temporal link between the descriptor and the acoustic signal that gave rise to it; for example, we don't know which vowels, in the above example, were perceived as defective; furthermore, we don't know what characteristics of the signal caused any segment to be perceived as a defective vowel.

Phonetic transcription, in that it links labels to short segments in the speech sequence, is a partial solution to this problem. The traditional technique within the speech pathology literature was to describe sounds as being "correct, substituted, omitted or distorted" (Templin, 1957). More recently, the refinement of feature description has been introduced (Compton, 1970; Pollack and Rees, 1972; Cairns, Cairns, and Williams, 1974; McReynolds and Engmann, 1975). Most of these analyses have been loosely based on the Chomsky and Halle (1968) treatment of the phoneme as a bundle of features, although the analyses differ in whether they consider only the features of the target phoneme, or the target and substituted phoneme, the feature system used for the analysis, and the status of distorted phones in the analysis. While these differences are substantial and affect the results of the analysis in major ways, the analysis procedure itself is perceptual, like the descriptor procedure, whether the feature labels have a physiological referent--like "voiced" or whether they do not--like "distributed." The characteristics of the speech perception mechanism itself are as much an issue in one of these descriptions as in another.

How good a guide is perceptual feature labeling to the underlying speech production? While there are sound theoretical reasons for believing in the interrelatedness of perception and production (as we discussed above), and in the perceptual reality of at least some features, at some level (Wickelgren, 1966; Studdert-Kennedy and Shankweiler, 1970), it can also be shown that listeners do not have a clear picture of the basis for their judgments.

In a recent paper, Strange, Verbrugge, Shankweiler, and Edman (1976) have shown that the identification of natural vowels is far better if the

vowels are produced in consonant-vowel-consonant (CVC) context than if they are produced in isolation. Apparently, then, transitional information is being used by the listener in making what appear as judgments of steady-state features, or targets. It is not, as the target formulations suggest, that a target vowel can be extracted from the changing signal; rather, the listener makes some use of the transition portion of the vowel in making the identification.

Listeners can be badly deceived as to what is wrong with defective speech. A good example is provided by a study of deaf speech, by Calvert (1961). He was examining the often reported observation that deaf speakers have a characteristic voice quality. He recorded deaf and normal speakers, producing simple sustained vowels and dissyllables. Although the listeners were experienced in listening to deaf speech, they could not discriminate between deaf and normal speakers unless they heard dissyllables; what was perceived as a voice defect was in fact an articulatory disorder.

The object of this discussion is not to condemn perceptual analysis of speech production. In fact, a number of investigators have made use of the types of analyses cited above, including the present author (Shankweiler and Harris, 1966). The problem is to make perceptual analyses yield better data about the speech production process. This is a two-stage problem: first, we need to find out more about the acoustic effects of movement disorders; second, using a manipulative approach, we need to find out more about what characteristics of the acoustic signal give rise to the perception of deviant speech.

#### IMPROVING THE USEFULNESS OF PERCEPTUAL ANALYSIS

Probably the most underused level of description of speech production is the acoustic signal. One way of getting a better understanding of the acoustic effects of deviant production is to synthesize it. Modern computer techniques make it possible to manipulate acoustic variables and study the perceptual effects. This approach amounts to an examination of the psychophysics of feature and descriptor perception. One might feel that the whole history of the study of the cues for speech perception represents such an attempt--for example, the study of the second formant transition as a cue for the voiced stops (Liberman, Delattre, Cooper, and Gerstman, 1954). However, an important limitation on these studies is that they were performed with total synthesis, that is, visual patterns were generated that seemed, on the basis of inspection of spectrograms, to provide likely cues for some discrimination between phones. Thus, one cannot tell whether a particular acoustic pattern was a possible output of the human articulatory system, or not. Neither does one know whether the cue manipulated in a given experiment is the one that a real speaker varies to generate the cue.

Another approach to the investigation of cues that has been tried in some laboratories is the manipulation of real speech. For example, a way of investigating intonation is to record and analyze real speech signals, change the fundamental frequency of the analyzed signal, and resynthesize.



This same kind of analysis by synthesis approach could be used with deviant speech. For example, in the example given of Calvert's study, the results appeared to show that deaf speakers were discriminated from normals because of deviant temporal patterns in their speech. This hypothesis could be checked experimentally by manipulating the durational characteristics of real samples.

In the case of Darley's descriptors, it would be possible to manipulate output deviant speech to isolate the factors giving rise to the judgment of deviancy, or to see what the "halo effect" of deviant articulation of one segment is on another.

I am not able to find many examples of this approach. Wendahl (1963; 1966) investigated perceived harshness by manipulation of a synthesized glottal tone. This was, however, a total synthesis experiment, and had the limitations described above.

I will conclude with a specific example, again from a study of deaf speech. It is well known that deaf speakers make a wide variety of errors and that the number of errors is highly correlated with the overall rated intelligibility of the speech. It can also be shown that children's speech shows gross temporal distortion, both because the words themselves are produced slowly, and because there are long pauses between words. Furthermore, formants for different vowels appear to be very similar, and do not show the usual amount of movement in time. A student, Mary Joe Osberger, is experimenting on this speech to see if its intelligibility will be improved by temporal manipulation. If the speech is unintelligible because it is slowly produced, then manipulation of the vowel durations, and elimination of the between-word pauses will make it intelligible. While it is instrumentally more complicated, the formant tracts can also be manipulated systematically by resynthesis.

This approach amounts to a kind of instrumental speech correction. The approach will, by inference, isolate the aspects of the acoustic signal judged to be deviant. Of course, this will still leave alternative articulatory mechanisms for generating the acoustic effect, as we can see by reference to the earlier sections of this paper. The importance of a perceptually based strategy, on the one hand, is that it keeps the physiological researcher from mere botanizing; on the other hand, it defines a class of physical variables that may, for the clinician, be deviant.

#### REFERENCES

- Abbs, J. H., J. W. Folkins, and M. Sivarajan. (1976) Motor impairment following blockade of the infraorbital nerve: implications for the use of anesthetization techniques in speech research. J. Speech Hearing Res. 19, 19-35.
- Amerman, J. D. and R. Danilooff. (1971) Articulation patterns resulting from modification of oral cavity size. ASHA, 13(9), S59(a).
- Angelocci, A. A., G. A. Kopp, and A. Holbrook. (1964) The vowel formants

- of deaf and normal-hearing eleven to fourteen year old boys. J. Speech Hearing Dis. 29, 156-170.
- Bell-Berti, F. (1976) An electromyographic study of velopharyngeal function in speech. J. Speech Hearing Res. 19, 225-240.
- Bell-Berti, F. and K. S. Harris. (1975) Some acoustic measures of anticipatory and carryover coarticulation. Haskins Laboratories Status Report on Speech Research SR-37/38, 73-78.
- Bell-Berti, F. and H. Hirose. (1975) Short communication. Palatal activity in voicing distinctions: a simultaneous fiberoptic and electromyographic study. J. Phonetics 3, 69-74.
- Benguerel, A.-P. and H. A. Cowan. (1974) Coarticulation of upper lip protrusion in French. Phonetica 30, 41-55.
- Bigland, B. and O. C. J. Lippold. (1954) The relation between force, velocity, and integrated electrical activity in human muscles. J. Physiol. 123, 214-224.
- Borden, G., K. S. Harris, and L. Catena. (1973) Oral feedback II. An electromyographic study of speech under nerve-block anesthesia. J. Phonetics 1, 297-308.
- Bouisset, S. (1973) EMG and muscle tone in normal motor activities. In New Developments in Electromyography and Clinical Neurophysiology, ed. by J. E. Desmedt, Vol. 1. (Basel: Karger).
- Cairns, H. S., C. E. Cairns, and F. Williams. (1974) Some theoretical considerations of articulation substitution phenomena. Lang. Sp. 17, 160-173.
- Calvert, D. R. (1961) Some acoustic characteristics of the speech of profoundly deaf individuals. Unpublished Ph.D. dissertation, Stanford University.
- Canter, G. J. (1963) Speech characteristics of patients with Parkinson's disease: I. Intensity, pitch, and duration. J. Sp. Hear. Dis. 28, 221-229.
- Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper and Row).
- Compton, A. (1970) Generative studies of children's phonological disorders. J. Sp. Hear. Dis. 35, 315-339.
- Daniloff, R. G. and R. E. Hammarberg. (1975) On defining coarticulation. J. Phonetics 1, 239-248.
- Daniloff, R. G. and K. L. Moll. (1968) Coarticulation of lip-rounding. J. Sp. Hear. Res. 11, 707-721.
- Darley, F. L., A. E. Aronson, and J. R. Brown. (1969a) Differential diagnostic patterns of dysarthria. J. Sp. Hear. Res. 12, 246-269.
- Darley, F. L., A. E. Aronson, and J. R. Brown. (1969b) Clusters of deviant speech dimensions in the dysarthrias. J. Sp. Hear. Res. 12, 462-496.
- Darley, F. L., A. E. Aronson, and J. R. Brown. (1975) Motor Speech Disorders. (Philadelphia: Saunders).
- Davis, S. B. (1976) Computer evaluation of laryngeal pathology based on inverse filtering of speech. Ph.D. Dissertation, U. Cal. at Santa Barbara.
- DeLuca, C. J. and W. J. Forrest. (1973) Probability distribution function of the interpulse intervals of single motor unit action potentials during isometric contractions. In New Developments in Electromyography and Clinical Neurophysiology, ed. by J. E. Desmedt, Vol. 1. (Basel: Karger).

- Dixit, R. and P. MacNeilage. (1972) Coarticulation of nasality: evidence from Hindi. J. Acoust. Soc. Amer. 52, 131(A).
- Fant, G. M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).
- Folkins, J. W. and J. H. Abbs. (1975) Lip and jaw motor control during speech: responses to resistive loading of the jaw. J. Sp. Hear. Res. 18, 207-221.
- Fritzell, B. (1969) The velopharyngeal muscles in speech: an electromyographic and cineradiographic study. Acta Otolaryngol., 250S.
- Gay, T. J. (1973) A cinefluorographic study of vowel production. J. Phonetics 2, 255-266.
- Hamlet, S. L. (1973) Speech adaptation to dental appliances: theoretical considerations. J. Baltimore College Dental Surgery 28, 52-63.
- Hamlet, S. L. and M. Stone. (in press) Reorganization of speech motor patterns following changes in oral morphology produced by dental appliance. In Proceedings of Speech Communication Seminar, Stockholm, Sweden, 1-3 August 1974. (Uppsala: Almqvist and Wiksell).
- Hamlet, S. L. and M. Stone. (1976) Compensatory vowel characteristics resulting from the presence of different types of experimental dental prostheses. J. Phonetics 4, 199-218.
- Harris, K. S., G. F. Lysaught, and M. M. Schvey. (1965) Some aspects of the production of oral and nasal labial stops. Lang. Speech. 8, 135-147.
- Hirose, H. (1975) The posterior cricoarytenoid as a speech muscle. Annual Bulletin of the Institute of Logopedics and Phoniatrics, No. 9, 47-66.
- Horii, Y., A. House, K. -P. Li, and R. Ringel. (1973) Acoustic characteristics of speech produced without oral sensation. J. Speech Hearing Res. 16, 67-77.
- Houde, R. A. (1967) A study of tongue body motion during selected speech sounds. SCRL Monograph No. 2. (Santa Barbara: Speech Communications Research Laboratory).
- Kent, R. D. (1972) Some considerations in the cinefluorographic analysis of tongue movements during speech. Phonetica 26, 16-32.
- Kent, R. D. and R. Netsell. (1975) A case study of an ataxic dysarthric: cineradiographic and spectrographic observations. J. Speech Hearing Dis. 40, 52-71.
- Kent, R. D., R. Netsell, and L. L. Bauer. (1975) Cineradiographic assessment of articulatory mobility in the dysarthrias. J. Speech Hearing Dis. 40, 467-480.
- Koike, Y. and J. Markel. (1975) Application of inverse filtering for detecting laryngeal pathology. Ann. Otol., Rhinol., Laryngol. 84, 117-124.
- Ladefoged, P. (1967) Three Areas of Experimental Phonetics. (London: Oxford University Press).
- Ladefoged, P., J. DeClerk, M. Lindau, and G. Papcun. (1972) An auditory-motor theory of speech production. Working Papers in Phonetics 22, (Linguistics Department U. Calif. at Los Angeles), 48-75.
- Leanderson, R., A. Persson, and S. Ohman. (1970) Electromyographic studies of the facial muscles in dysarthria. Acta Otolaryngol. 263, 89-94.
- Lehiste, J. (1965) Some acoustic characteristics of dysarthric speech. Biblioteka Phonetica Fasc. 2. (Basel: Karger).
- Lieberman, A. M., F. S. Cooper, K. S. Harris, and P. F. MacNeilage.

- (1963). A motor theory of speech perception. In Proceedings of the Speech Communication Seminar, Stockholm, 1962. (Stockholm: Royal Institute of Technology).
- Lieberman, A. M., P. C. Delattre, F. S. Cooper, and L. J. Gerstman. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr. 68, no. 379.
- Lieberman, P. (1963) Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. J. Acoust. Soc. Amer. 35, 344-353.
- Lieberman, P. (1973) On the evolution of language: a unified view. Cognition 2, 59-94.
- Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773-1781.
- Lindblom, B., J. F. Lubker, and B. Fritzell. (1974) Experimentalfonetiska studier av dysartri. Papers from the Institute of Linguistics, University of Stockholm 27.
- Lindblom, B. and J. E. F. Sundberg. (1971) Acoustical consequences of lip, tongue, jaw and larynx movement. J. Acoust. Soc. Amer. 50, 1166-1179.
- Lisker, L. (1975) Of time and timing in speech. In Current Trends in Linguistics, vol. 12, ed. by T. Sebeok. (The Hague: Mouton).
- Lubker, J. F., B. Fritzell, and J. Lindqvist. (1970) Velopharyngeal function: an electromyographic study. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm), STL-QPSR 4, 9-20.
- MacNeilage, P. F. (1963) Electromyographic and acoustic study of the production of certain clusters. J. Acoust. Soc. Amer. 35, 461-463.
- MacNeilage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev. 77, 182-195.
- MacNeilage, P. F. (1973) Preliminaries to the study of single motor unit activity in speech musculature. J. Phonetics 1, 55-71.
- MacNeilage, P. F. and J. DeClerk. (1969) On the motor control of coarticulation in CVC monosyllables. J. Acoust. Soc. Amer. 45, 1217-1233.
- MacNeilage, P. F., H. M. Sussman, and R. K. Powers. (in press) Firing patterns of single motor units in speech musculature. Proceedings of the 8th International Congress of Phonetic Sciences, Leeds, England, 17-23 August 1975.
- Markel, J. D. and A. H. Gray. (1976) Linear Prediction of Speech. (Berlin: Springer Verlag).
- McCall, G. N., M. L. Skolnick, and D. W. Brewer. (1971) A preliminary report of some atypical patterns in the tongue, palate, hypopharynx and larynx of patients with spasmodic dysphonia. J. Speech Hearing Dis. 36, 446-470.
- McReynolds, L. V. and D. L. Engmann. (1975) Distinctive Feature Analysis of Misarticulations. (Baltimore: University Park Press).
- Moll, K. L. (1960) Cinfluorographic techniques in speech research. J. Speech Hearing Res. 3, 227-241.
- Monson, R. B. (1974) Durational aspects of vowel production in the speech of deaf children. J. Speech and Hearing Res. 17, 386-398.
- Monsen, R. B. (1976) Second formant transitions of selected consonant-vowel combinations in the speech of deaf and normal-hearing children. J. Speech Hearing Dis. 19, 279-289.
- Moore, P. (1938) Motion picture studies of the vocal folds and vocal



- attack. J. Speech Hearing Dis. 3, 235-238.
- Netsell, R. and C. S. Cleeland. (1973) Modification of lip hypertonia in dysarthria using EMG feedback. J. Speech Hearing Dis. 38, 131-140.
- Netsell, R., B. Daniel, and G. C. Celestia. (1975) Acceleration and weakness in Parkinsonia dysarthria. J. Speech Hearing Dis. 40, 170-178.
- Netsell, R. and R. D. Kent. (1976) Paroxysmal ataxic dysarthria. J. Speech Hearing Dis. 41, 93-109.
- Nooteboom, S. G. (1970) The target theory of speech production. IPO Annual Progress Report 5, (Institute for Perception Research, Eindhoven, Holland), 51-55.
- Ohde, R. N. and D. J. Sharf. (1975) Coarticulatory effects of voiced stops on the reduction of acoustic vowel targets. J. Acoust. Soc. Amer. 58, 923-927.
- Ohman, S. E. G. (1965) Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Amer. 39, 151-168.
- Pollack, E. and N. Rees. (1972) Disorders of articulation: some clinical applications of distinctive feature theory. J. Speech Hearing Dis. 37, 451-462.
- Putnam, A. H. B. and R. L. Ringel. (1972) Some observations of articulation during labial sensory deprivation. J. Speech Hearing Res. 15, 529-542.
- Ringel, R. L. and M. D. Steer. (1963) Some effects of tactile and auditory alterations on speech output. J. Speech Hearing Res. 6, 369-378.
- Rothenberg, M. (1973) A new inverse filtering technique for deriving the glottal air flow waveform during voicing. J. Acoust. Soc. Amer. 53, 1632-1645.
- Sawashima, M. and H. Hirose. (1968) New laryngoscopic technique by use of fiberoptics. J. Acoust. Soc. Amer. 43, 168-169.
- Scott, C. M. and R. L. Ringel. (1971a) Articulation without oral sensory control. J. Speech Hearing Res. 14, 804-818.
- Scott, C. M. and R. L. Ringel. (1971b) The effects of motor and sensory disruption on speech: a description of articulation. J. Speech Hearing Res. 14, 819-828.
- Shankweiler, D. and K. S. Harris. (1966) An experimental approach to the problem of articulation in aphasia. Cortex 2, 277-292.
- Skolnick, M. L., G. N. McCall, and M. Barnes. (1973) The sphincteric mechanism of velopharyngeal closure. Cleft Palate J. 10, 286-305.
- Smith, T. and C. Lee. (1972) Peripheral feedback mechanisms in speech production models. In Proceedings of the 7th International Congress of Phonetic Sciences, Montreal, 1971, ed. by A. Rigault and R. Charbonneau. (The Hague: Mouton).
- Stevens, K. N. (1973) The quantal nature of speech: evidence from articulatory-acoustic data. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw Hill).
- Stevens, K. N. and A. S. House. (1955) Development of a quantitative theory of vowel articulation. J. Acoust. Soc. Amer. 27, 484-493.
- Stevens, K. N. and A. S. House. (1963) Perturbations of vowel articulations by consonantal context: an acoustical study. J. Speech Hearing Res. 6, 111-128.
- Stevens, K. N., A. S. House, and A. P. Paul. (1966) Acoustic description of syllable nuclei: an interpretation in terms of a dynamic model of articulation. J. Acoust. Soc. Amer. 40, 123-132.
- Strange, W., R. R. Verbrugge, D. P. Shankweiler, and T. R. Edman. (1976).

- Consonant environment specifies vowel identity. J. Acoust. Soc. Amer. 60, 213-224.
- Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.
- Subtelny, J. D., H. Koepp-Baker, and J. D. Subtelny. (1961) Palatal function and cleft palate speech. J. Speech Hearing Dis. 26, 213-224.
- Templin, M. (1957) Certain Language Skills of Children. (Minneapolis: U. Minnesota Press).
- Ushijima, T. and H. Hirose. (1974) Electromyographic study of the velum during speech. J. Phonetics 2, 315-326.
- von Leden, H., P. Moore, and R. Timke. (1960) Laryngeal vibrations: measurements of the glottic wave. part 3, The pathologic larynx. Archives Otolaryngol. 71, 16-35.
- Wendahl, R. W. (1963) Laryngeal analog synthesis of harsh voice quality. Folia Phoniat. 15, 241-250.
- Wendahl, R. W. (1966) Laryngeal analog synthesis of jitter and shimmer, auditory parameter of harshness. Folia Phoniat. 18, 98-108.
- Wickelgren, D. (1966) Distinctive features and short-term memory for English consonants. J. Acoust. Soc. Amer. 39, 388-398.

# Universals in Phonetic Structure and Their Role in Linguistic Communication\*

Michael Studdert-Kennedy†

## ABSTRACT

All known languages display duality of patterning, phonological system and structure. All spoken languages are syllabic and constrain perceptual structure in terms of consonants and vowels. The syllable is a unit of timing in articulation, of contrast and compression in perception. These functions arise from the temporal structure of an acoustic signaling system and are fulfilled by spatial structure in the visual signaling system of American Sign Language. Syllabic compression poses a problem for the perceiver, if he is to recover discrete message units from an unsegmented signal. A possible mechanism based on acoustic feature detectors, for accomplishing the segmentation, and an alternative process, based on perceptual contrast and continuous tracking of the signal, are considered.

## INTRODUCTION

If we are to compare the acoustic signaling systems of humans and other animals, we must begin by distinguishing between the physical signal and the perceived message. To understand the importance of this distinction, compare the approaches of a phonetician and a cryptographer to the spectrographic display of a spoken utterance. The phonetician comes armed with preconceptions as to how the utterance is to be segmented--into phonemes, syllables, words--and, naturally enough, seeks in the spectrogram the acoustic correlates of those segments. The fact that he finds, on the one hand, more segments than he wants and, on the other hand, many acoustically indivisible segments that he knows must correspond to two or more perceived segments, has been a main reason for the hypothesis that speech perception engages specialized decoding mechanisms. By contrast, the cryptographer, as the student of birdsong, will begin by dividing the utterance into acoustically distinct segments. Then, will all the means at his disposal--systematic distributional analysis of other utterances, segment transposition, informant tests, and so on--he will try to determine the groupings and cuts in the acoustic signal necessary to derive the set of functional (as opposed to physical) segments that constitute the message. Not surprisingly, phoneti-

---

\*This paper was presented at the Dahlem Konferenzen, Berlin, 27 Sept -1 Oct., 1976 and will be published in the conference proceedings Dahlem Konferenzen, ed. by T. H. Bullock.

†Also Queens College and Graduate Center of the City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

cians have shrunk from this task. They have preferred to accept the segments of abstract linguistic analysis and to search for their correlates in the signal. For the present, we have little choice but to continue this tradition, and most of what I have to say circles around the resulting problem of segmentation. However, before I come to this, a few further general points must be made.

#### SOME LANGUAGE UNIVERSALS

All known languages (and perhaps some animal communication systems) display "duality of patterning" (Harris, in press). Utterances can be described, at the syntactic level of the message, as sequences of lexical and grammatical segments (words or morphemes), and at the phonological level, as sequences of meaningless segments (phonemes). Words leave their faint traces only in prosodic features of the signal, if at all, and our present concern is entirely with the "lower" meaningless segments.

All known spoken languages have a sound system, or phonology, based on feature opposition. Sound units, the phonemes of a language, are a relatively small set (usually a few dozen) of meaningless segments that serve to distinguish among its words. For example, the words "bed" and "red" are distinguished by their initial, but not by their medial or final phonemes. The phonemes are not randomly selected. Each has a characteristic internal structure that may be described in terms of the small set of phonetic features (usually a dozen or so) used in a particular language. The phonemes may be classified according to their shared phonetic features, and the resulting classes contrasted with one another on the basis of their feature differences or oppositions. For example, English /b/ and /p/, formed by closure of the vocal tract at the lips, share the feature "labial". They contrast with /d/ and /t/, formed by closure of the tract at the gum ridge behind the upper teeth, and termed "alveolar". At the same time, /b/ and /d/ share the phonetic feature "voiced" and contrast with "voiceless" /p/ and /t/. Taken together, these four phonemes constitute a little system of feature similarities and oppositions, such that /b/:/d/ = /p/:/t/ and /b/:/p/ = /d/:/t/. Phonemic, or feature, oppositions are presumed to be reflected in the signal by acoustic contrast.

All known spoken languages display phonological structure, that is, they constrain the arrangements in which phonemes may be combined to form words. These constraints reflect and, in part, define the feature classifications within the system: the domain of a phonotactic rule is a feature class. For example, in English, a stop consonant following initial /s/ is always voiceless (as in "spy", "sty" or "sky"); or, initial voiced stop consonants cannot be followed by nasal consonants (/bn-/ or /dn-/ for example, is not permitted). Phonological constraints of this kind are reflected in the signal by the types of sequential acoustic contrast permitted in a language.

I have rehearsed these generalities in order to raise the question of form and function. The particular form taken by the phonology of a language results from complex historical and social forces, as well as from phonetic, syntactic and semantic forces within the language itself. The general form taken by the phonologies of all languages, that is, the phonetics of language must reflect, at least in part, the constraints of physiology and communica-



tive efficiency: its boundary conditions are set by what we can articulate and what we can perceive.

Several likely preconditions of linguistic communication suggest plausible functions for the "lower" level of the dual pattern: (1) learnability: a limited set of feature oppositions and a limited set of permissible sequences must facilitate language acquisition and retention; (2) lexical productivity: a system of minimal sound opposition between meaningful segments increases potential semantic range and flexibility; (3) memorability: both long-term and short-term memory must be facilitated. The ability to store meaningless phonemic sequences, pending syntactic and semantic processing, as we listen and perhaps as we speak, may be a condition of complex syntactic processing. The form of this store is a matter of considerable interest and has been the target of many short-term memory studies [see Studdert-Kennedy (1976) for a review].

All that I have said so far has been intended to draw attention to the crucial role that segmentation, whether by feature or by phoneme, plays in language and speech, and to set the stage for discussion of a structure common to all languages, the consonant-vowel syllable.

#### A UNIVERSAL OF PHONETIC STRUCTURE: THE CONSONANT-VOWEL SYLLABLE

All spoken languages are syllabic. All languages constrain syllable structure in terms of consonants and vowels. All languages permit the consonant-vowel (CV) syllable. For all languages this is the canonical speech gesture.

#### Articulatory Function of the Syllable

The reason why all spoken languages are syllabic is not hard to find. Nothing is easier than to open the mouth: the CV syllable is fundamentally an articulatory unit. The consonant-vowel feature opposition is between a constricted and an open vocal tract, the two shapes being combined in a single "ballistic gesture" (Stetson, 1951:4). The articulatory function of the syllable is probably as a unit of neural timing in the control of speech (Fry, 1964; Kozhevnikov and Christovich, 1965). Control of the speech musculature--of the muscles for breathing, phonating and articulating--is obviously a very complicated affair, requiring precise coordination of its parts. The time window over which coordination is accomplished might, in principle, be fixed, with new specifications appearing automatically as earlier specifications are implemented. The process would then be continuous and unsegmented, with new specifications adjustable to earlier, but not to later, commands. However, studies of speech production have shown that both anticipatory and perseverative coarticulation occur (Harris, in press). For this to be so, two or more phonetic segments must be programmed simultaneously. The decision as to how many units will be programmed at a time may then be determined at the lowest level by the composition of the syllable (which in many languages may include several consonants). Although no detailed model of the articulatory process has been worked out, this view of the syllable is not incompatible with the fact that coarticulation occurs across syllable boundaries, since this may be presumed to arise at a higher level of

temporal coordination. That several levels of coordination may be required is suggested by the precision with which speech rhythm can be controlled over a lengthy utterance, despite considerable internal variation in segment duration.

#### Perceptual Function of the Syllable

The acoustic consequences of syllabic gestures can be observed in the undulations of an oscillogram. The perceptual function of these variations is to carry contrasts in stress, rhythm, and intonation. None of these contrasts would be possible without the vowel, their carrier. Yet a language that consisted entirely of vowels would soon exceed the acceptable limits of homonymy. The addition of various forms of vowel onset or "attack" (that is, of consonants) to the phonetic repertoire provides the acoustic ground of perceptual contrast and releases the potential of dual patterning.

A second consequence of the syllabic gesture is evinced in the standard spectrogram by the apparent absence of acoustic segmentation within many stretches of the signal known to correlate with syllables. The effect of this compression is to increase the rate of signaling phonetic segments (or phonemes). Precise limits on signaling rate are not known. However, on the perceptual side, an upper limit must be set by the ability of the ear to resolve a succession of more or less discrete segments, and a lower limit must be set by our short-term memory for unparsed stretches of speech. If the lower limit exceeds the upper, a possible solution is to merge units into larger units and so reduce the segment rate. This is exactly what is accomplished by the syllable.

However, syllabic compression has the apparent disadvantages of destroying the acoustic basis for recovering the phonetic segments of which dual patterning is premised. The difficulty is exacerbated by the spread across an entire syllable of acoustic properties presumed to characterize only one of the component phonetic segments. This effect, although often treated separately as a manifestation of "the invariance problem", is equally puzzling for an account of segmentation: to segment a syllable we must not only separate the pieces that do not belong together, but we must also connect those that do. Before examining this problem in more detail, let us consider how syllabic functions are fulfilled in a language that uses another modality.

#### SOME COMPARISONS WITH AMERICAN SIGN LANGUAGE

In recent years several laboratories in the United States have been intensively investigating American Sign Language (ASL) (Hockett, 1958; Klima and Bellugi, in press). ASL is the language of many thousands of deaf persons in the United States. Its medium is manual (primarily), facial, and bodily gesture. Here the arbitrary structure of ASL signs is of particular interest. Each sign is a meaningful unit, translatable into one or more English words, but signs are neither perceived nor remembered as "wholes". Intrusion errors in short-term memory reflect formational properties of the signs, analogous to the phonological intrusions of reading and hearing studies (Bellugi and Klima, 1975). Manual signs vary along the dimensions of shape, location, orientation and movement. Within each dimension there

appears to be a limited set of oppositions. Although it is not yet clear precisely what these oppositions are, clustering and scaling analyses of errors in the perception of hand shapes in visual noise (Lane, Boyes-Graem and Bellugi, 1976) have achieved at least the quantitative validity of comparable speech studies (Miller and Nicely, 1955). Furthermore, permissible feature combinations are a small set of the possible combinations, and rules for such standard formational ("phonological") processes as deletion and assimilation have already been described (Battison, 1974; Frishberg, 1975). It is probably a mere matter of time before a "phonology" of ASL is derived, and before work on the "universal phonetics" of sign language begins. What we have then in ASL (as, no doubt, in the sign languages of China, England, France, and many other communities) is a system of communication that displays syntax, formational system and formational structure: in short, duality of patterning. How, we may ask, are the syllabic functions of contrast and compression fulfilled in this language?

Sign uses a spatial rather than a temporal medium. Although signs may differ in movement, they are primarily distinguished by simultaneous, not sequential contrasts. They are linked sequentially to form "utterances", so that coarticulation over time does occur (for example, a hand may adopt the shape of a following sign before the preceding sign has been completed). But temporal coarticulation is not intrinsic to sign as it is to speech, because its units are units of spatial rather than temporal coordination. The functions of contrast and compression are therefore fulfilled simultaneously. The analog of the syllable is the sign itself.

We may point up the analogy, and the difference, by noting that the feature oppositions of both speech and sign are articulatory, and that fine motor control is typically vested in the same cerebral hemisphere for both dominant hand and mouth. [One hand is designated the "more active" in ASL formational rules (Battison, 1974), but which one depends on the handedness of the signer.] We are then led to wonder whether speech, viewed through transparent skin rather than heard as the acoustic consequences of its articulation, could be "read" as directly as sign. Probably it could not: for it is precisely in the temporally organized gestures of speech that the phonetic segments are lost. The segmentation problem may therefore be peculiar to speech.

#### THE ROLE OF ACOUSTIC FEATURE DETECTORS

One approach to the problem is implicit in the work of Stevens (Stevens, 1972; Stevens, 1975) who has argued "...that there is some justification on a purely physical basis for characterization of phonemes in terms of discrete properties or features." (Stevens, 1972:53). He shows that there are configurations of the vocal tract for which relatively large changes in articulation lead to relatively small changes in acoustic output. These regions are bounded by others for which precisely the reverse relation holds: small changes in articulation lead to large changes in acoustic output. He illustrates this principle with instances of articulatory-acoustic distinctions among important distinctive feature classes used in many languages, and suggests that the phonetic inventory of all languages is assembled from these "quantal" regions.

At first sight, one might take Stevens to be offering a straight-forward description of articulatory-acoustic parameters that could be used for speech synthesis or automatic speech recognition. However, he also states that a requirement for selection of quantal regions for phonetic use is "...that the attributes of the signal be relatively insensitive to articulatory perturbations after the signal is transformed by the auditory mechanism" [Stevens, (1972:64), *italics in the original*]. In other words, not only must the acoustic signal be relatively insensitive to articulatory perturbation, but the auditory system must be relatively insensitive to acoustic perturbation: it must perceive categorically. To meet this requirement, Stevens, in a later paper (Stevens, 1975), posits the existence of "auditory property detectors", tuned (or tunable) to the acoustic properties of speech. One explicitly stated reason for positing these detectors is that they provide a mechanism by which the infant might latch onto the phonetically relevant properties of speech. A second reason not mentioned by Stevens but, if I understand him, essential to his account, is that they might segment the flow of speech. For it matters little that speech is "quantal", if we have no device for sifting the quantal properties from the flow. In short, far from discovering the message in the signal, as it were, Stevens is offering an explicitly physiological account, to which the existence of discrete property detectors is essential.

A great deal of research in the past few years has been directed toward isolating such detectors. Eimas (Eimas and Corbit, 1973) introduced an "adaptation" procedure, modified from visual research. He and his colleagues showed that repeated exposure to a stimulus possessing a particular acoustic feature (for example, the rising formant transitions of some labial stop consonants) reduced a subject's sensitivity to that feature and relatively increased his sensitivity to an opponent feature (the falling transitions of some apical stop consonants). The procedure has been effectively used with many types of speech stimuli, contrasting in minimal acoustic features known to be associated with phonetic feature oppositions. The results have generally been interpreted as evidence for detectors tuned to the manipulated feature.

Let us suppose that this interpretation is correct and that banks of acoustic feature detectors, or analyzers, are neatly sprung by the syllabic flow. What then has been gained? First, a degree of segmentation. Second, a recoding of the signal from continuous to discrete form, so as to allow short-term storage of information without "echoic" decay. These are important gains, but they do not move us very far toward a phonetic interpretation of the signal. One reason for this is that they can, at best, perform one half of the segmentation: they may separate what must be separated, but they cannot connect what must be connected.

Consider, for example, how feature detectors would analyze the existential injunction, "Be!" (/bi/). The following features might be detected: (1) silence, (2) a rapid upward shift of the spectrum, voiceless for, say, 10 msec, voiced for 30 msec, in the vicinity of the second and third formants, (3) a brief delay (10 msec) between the onset of the spectral shift and the onset of glottal pulsation, (4) a rapid (30 msec), small upward shift of the first formant at the onset of glottal pulsation, and (5) a relatively sustained formant pattern. The phonetician knows that the first four



features are typical of voiced labial stops before high front vowels. However, the cryptographer (that is to say, the auditory system) does not. What auditory principle groups the first feature, silence, with the next three, but fails to group the fourth with the fifth? In other words, what auditory principle integrates the acoustic features of the consonant and separates them from those of the vowel?

If we must rely on feature analyzing systems, there seems, in fact, to be none. The proposed detectors thus lead us into an impasse from which we can only escape by invoking some nonauditory principle of perceptual organization--precisely the impasse they were intended to avoid.

#### PERCEPTUAL CONTRAST AND CONTINUOUS TRACKING

The source of the difficulty is the desire to match our percepts with both phonological description and the acoustic signal. Perhaps we have been misled in attempting to model perceptual performance after the linguist's model of phonological competence. We cannot evade the dual pattern. Must the perceptual segments be static? Why, if speech is acoustic and if the essence of an acoustic event is its temporal organization, are features and phonemes commonly defined as points in space, static configurations of the vocal tract, or as stationary auditory qualities?

If we return to the canonical speech gesture, two facts stand out. First, the articulatory poles of the syllable--constricted vs. open--provide maximal perceptual contrast. Second, the contrast is always and only manifested over time. The syllable is a unitary event of which the auditory quality (or phonetic manner) changes as it occurs. If we perceived the contrast directly, as a development, much as we perceive the attack and sustention of a musical note, without benefit of specialized detectors to "stop the image", the contrast would be the ground of our perceptual segmentation.

Our percepts would not then be segments, but acoustic events for which we happen to have a segmental notation (arrayed in space). From this point of view, the perceptual process in a continuous tracking of an acoustic signal, isomorphic, point for point, with the continuously changing articulation. The perceptual elements of the dual pattern would not then be the timeless entities of current phonology, but dynamic events, jointly shaped by the timing mechanisms of motor control and by the demands of the auditory system for perceptual contrast and compression.

#### REFERENCES

- Battison, R. (1974) Phonological deletion in American Sign Language. Sign Language Studies 5, 1-19.
- Bellugi, U., E. S. Klima, and P. A. Siple. (1975) Remembering in signs. Cognition 3, 93-125.
- Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation and linguistic feature detectors. Cog. Psych. 4, 99-109.
- Erishberg, N. (1975) Arbitrariness and iconicity: historical change in American Sign Language. Language 51, 696-719.
- Fry, D. B. (1964) The function of the syllable. Z. Phon., Sprachwiss.

- Komm. Fschg. 17, 215-221.
- Harris, K. S. (in press) The study of articulatory organization: Some negative progress. In Research on Dynamics of Speech Production, Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, Tokyo, Japan.
- Hockett, C. F. (1958) A Course in Modern Linguistics. (New York: MacMillan).
- Klima, E. S. and U. Bellugi. (in press) The Signs of Language. (Cambridge: Harvard University Press).
- Kozhevnikov, V. A. and L. A. Chistovich. (1965) Rech' Artikuliatsia i vospriiatie. Moscow, Leningrad. Translated as Speech: Articulation and Perception. (Washington: Clearinghouse for Federal Scientific and Technical Information, J.P.R.S.), 30.
- Lane, H., P. Boyes-Graem, and U. Bellugi. (1976) Preliminaries to a distinctive feature analysis of hand shapes in American Sign Language. Cog. Psych. 8, 263-289.
- Miller, G. A. and P. E. Nicely. (1955) An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27, 338-352.
- Stetson, R. H. (1951) Motor Phonetics. (Amsterdam: North-Holland).
- Stevens, K. N. (1972) The quantal nature of speech: evidence from articulatory-acoustic data. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw Hill), 51-66.
- Stevens, K. N. (1975) The potential role of property detectors in the perception of consonants. In Auditory Analysis and Perception of Speech, ed. by C. G. M. Fant and M. A. A. Tatham. (New York: Academic Press), 303-330.
- Studdert-Kennedy, M. (1976) Speech Perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press), 243-293.

# Difference Limens for Formant Frequencies for Steady-State and Consonant-Bounded Vowels\*

Paul Mermelstein and Hollis Fitch†

## ABSTRACT

Difference limens (DL) of formant frequencies were measured for two steady-state vowels and the same vowels in symmetric stop-consonant contexts. The stimuli were generated using a computer-programmed synthesizer and the formant-frequency parameters were adjusted to be steady or symmetric cubic functions of the time difference from the temporal center of the syllable. The DL for the time-varying consonant-vowel-consonant (CVC) stimuli were found to be significantly larger than those for the steady-state vowels. In some cases the DL for the second formant is found to be larger in the direction of expected formant shift due to consonantal coarticulation, than in the reverse direction. The difference in DL values in and out of context has, at least partially, an auditory origin. However, the phonetic decoding of the CVC stimuli may introduce additional information loss from auditory short-term memory.

## INTRODUCTION

Difference limens (DL) for steady-state vowels reflect the ability of the whole auditory system to differentiate complex stimuli with stationary spectral patterns. The speech signal, however, is rarely stationary in its spectral composition for any length of time. The time-varying formant patterns form significant cues for the extraction of phonetic information. It is of interest, therefore, to explore the effects of spectral variation on the just noticeable differences (JND) in formant frequencies. The JNDs measured in consonantal context can be expected to be better indicators of formant-frequency discrimination in continuous speech.

It has been previously reported that the perceived color of vowels presented in isolation varies in a continuous manner. Discrimination of changes in vowel quality is equally acute within a phoneme region or across a phoneme boundary (Fry, Abramson, Eimas, and Liberman, 1962). Yet vowels in context tend to be perceived in a categorical fashion, that is, they possess discrimination functions that are characterized by peaks at the phoneme boundaries (Stevens, 1968).

---

\*Presented at the 92nd meeting of the Acoustical Society of America, San Diego, Calif., 16-19 November 1976.

†University of Connecticut, Storrs.

The effects of context are to introduce different time-dependent formant variations in the vocalic segments. In the presence of such variations, the vowel-category boundaries are displaced as well (Lindblom and Studdert-Kennedy, 1967). Acoustic data on vowel reduction (Stevens and House, 1963) reveal that the characteristic formant frequencies of vowels in isolation are not attained in dynamic speech contexts. It has been suggested that during the perception of vowels in consonantal contexts a corresponding compensation is made for undershoot in vowel articulation. (Lindblom and Studdert-Kennedy, 1967; Stevens, 1968).

Speech coding systems extract a set of time-varying parameters from the speech signal at the source, transmit the values of these parameters, and reconstitute the signal at the destination. The DLs of these parameters represent the maximum permissible transmission errors if the signal is not to be noticeably degraded in transmission. The bandwidth requirements of the encoded signal are therefore directly dependent on the DLs of the parameters. Any significant increase in the DL as the size of the context is increased would suggest that by selecting such larger signal segments for encoding as a unit, bandwidth savings in the transmission of the individual parameters may be attained beyond those due to the inherent limitations on the time variations of these parameters.

Measurement of DL attempts to exclude explicit labeling of the speech stimuli from the perception process by focusing on discriminable differences irrespective of their origin. Identification of stimuli must be based on discriminable differences. Thus, not only are actual DL values of interest in designing communication systems (Flanagan, 1955), but also as indicators of human speech perception performance. We have been particularly interested in the relative DL values at various points in the formant-space for vowels in and out of context.

Strange, Verbrugge, Shankweiler, and Edman (1976) report better identification of vowels in a consonantal environment than when spoken in isolation. These results are interpreted as evidence that dynamic acoustic information distributed over the temporal course of the syllable is used by the listener to identify vowels. If consonant-embedded stimuli vary in their steady-state formant frequency values but not in the initial and final frequencies, they must also vary in the transitional segments. If this transitional information is particularly useful for vowel discrimination, then listeners should discriminate synthesized consonant-embedded vowels better than steady-state vowels similarly synthesized. If such improved discrimination is not found, the improved identification of naturally spoken vowels in consonant-vowel-consonant context over isolated vowels must depend on other factors.

Stevens (1968) reported higher discrimination functions for isolated vowels than for CVC stimuli. The CVC stimuli used in his study, however, varied both in steady-state formant frequency and duration. The improved discrimination may have been due to the additional temporal differences among the CVC stimuli. Our study focuses on discrimination in formant frequencies alone and forms an extension of Stevens' results to different vowels and contexts.



## Stimuli

Steady-state vowels and CVC syllables were generated using a software formant-synthesizer modeled after the one presented by Rabiner (1968). The first three formants were adjustable under program control; the fourth and fifth formants were fixed at 3500 and 4500 Hz respectively. Bandwidth values were fixed at 60, 80, 100, 175, and 281 Hz respectively. The stimuli were entirely voiced and consisted of steady or changing formant-patterns. The formant trajectories for the CVC syllables followed symmetric cubic functions of the time differences from the temporal center of the stimulus.

The paired stimuli consisted of a standard and a variable counterbalanced to eliminate any order effects. The variable stimuli had central formant frequency values incremented in  $F_1$  (+ 25 Hz steps) or  $F_2$  (+50 Hz steps). Differences of one, two, three, and four steps were used in each increment direction to cover the expected DL range as judged from the data given by Flanagan (1955). Two symmetric consonantal contexts were used--/bVb/ and /gVg/. These differed only in trajectory of the second formant frequency. All of the stimuli were 200 msec in duration except for an additional shorter steady-state series of 133 msec duration that is discussed further below. The formant frequency variations are illustrated in Figure 1. The steady-state vowels had a steady fundamental frequency of 125 Hz except for a drop to 100 Hz over the last 50 msec. The fundamental frequency of the CVC stimuli followed the same cubic trajectory as the formant frequencies and covered the same range, namely 125 Hz at the center and 100 Hz at the initial and terminal points. The different fundamental frequency variations chosen for the CVC and V stimuli contributed somewhat to an enhanced naturalness for each group.

In view of Stevens' (1968) results that vowels in consonantal context near the phoneme boundary are discriminated better than vowels with formant values well within the phoneme-category, we used two separate standard stimuli as given in Table 1. In the first experiment, we used values appropriate for the vowel /i/, in the second, a value near the boundary between /ε/ and /æ/ (Peterson and Barney, 1952). Through use of different standards, we attempted to explore what effects, if any, proximity to a phoneme boundary may have on the DL values.

TABLE 1: Central and terminal formant frequencies for the standard stimuli.

|               |                 | Formant frequencies (Hz) |     |           |
|---------------|-----------------|--------------------------|-----|-----------|
| Experiment I  | Vowel /i/       | (center)                 | 350 | 2100 2900 |
|               | Stop /b/        | (terminal)               | 50  | 1500 2000 |
|               | Stop /g/        | (terminal)               | 50  | 2400 2600 |
| Experiment II | Vowel /ε/ - /æ/ | (center)                 | 600 | 1780 2450 |
|               | Stop /b/        | (terminal)               | 50  | 900 2000  |
|               | Stop /g/        | (terminal)               | 50  | 2100 2300 |

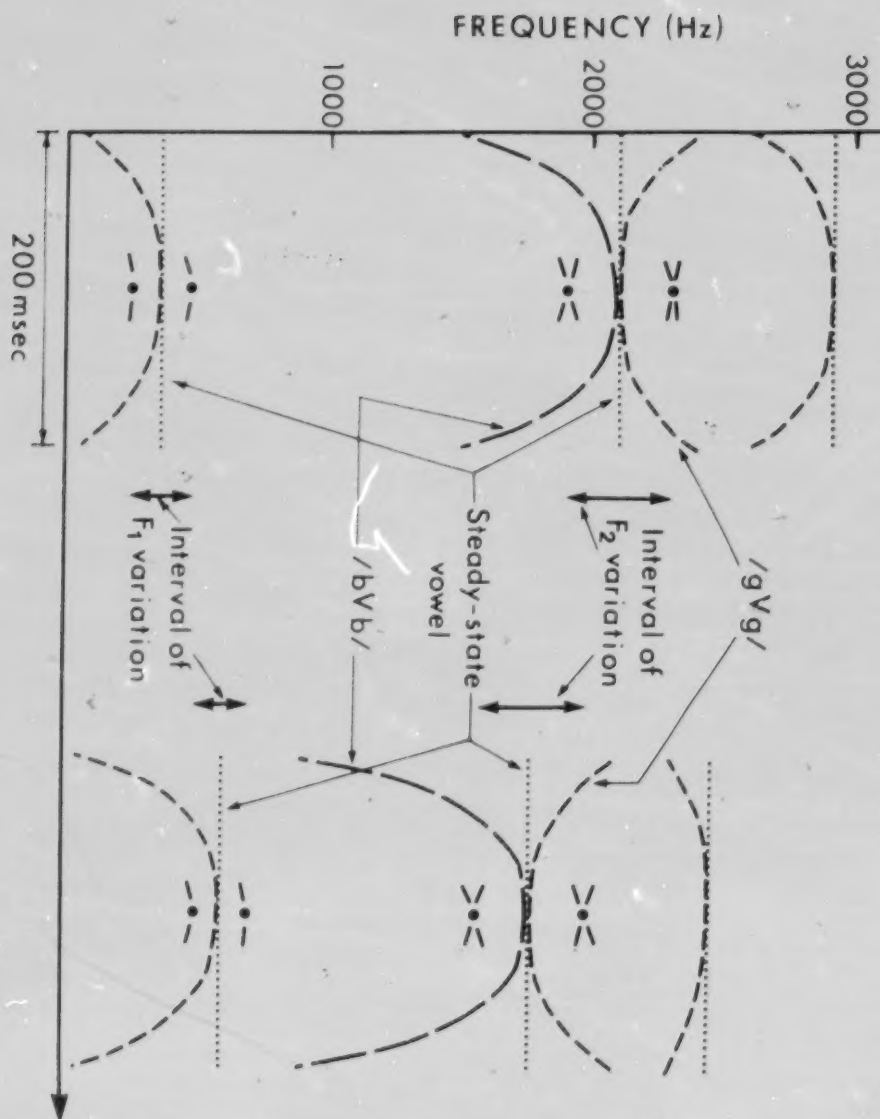


Figure 1: Formant trajectories for steady-state vowels and consonant-vowel-consonant stimuli. Left-[r], [brb], [grg] series; right-[ε], [bεb], [gεg] series. Unconnected points at syllable centers indicate ranges of  $F_1$  and  $F_2$  variation.

## Subjects

The subjects were volunteer students from Yale University and the University of Connecticut, and were paid \$2.00 per hour for their services. Five subjects took part in the first experiment, and six subjects participated in the second experiment. An additional group of five subjects, colleagues and students at Haskins Laboratories, took part in the identification experiments. All subjects were native speakers of English.

## Procedure

Pisoni (1971) reports that lower scores are obtained for steady-state vowel discrimination with the ABX technique than with other procedures. He suggests that the heavy memory-related task requirement associated with the ABX may conceal differences among stimuli that are mainly due to acoustic factors. Fujisaki and Kawashima (1970) model the decision process for ABX discrimination as a two-stage process. If A and B lie on opposite sides of a phonetic boundary, then the listener must determine only whether X belongs to the same category as A and B. However, if A and B are determined to belong to the same phonetic category, then X must be compared with the auditory memories of A and B. To avoid such memory-related problems, we attempted to simplify the discrimination task by using paired stimuli and asking for a "same/different" response.

Subjects heard pairs of stimuli 0.4 seconds apart, separated by 1.6 secs of silence from the next pair. They were asked to judge whether each pair consisted of the same or different stimuli. Some pairs consisted of the standard presented twice; others consisted of the standard and variable in either order. Steady-state vowel pairs, and /bVb/ and /gVg/ syllable-pairs were all randomized within one grand list. Each pair occurred at least five times each order.

Subjects were allowed to listen to trial pairs of stimuli until they felt comfortable with the quality of the synthetic speech. Roughly 10 pairs usually sufficed. Each experiment lasted approximately one-half hour and was divided into two parts by a short rest period. The first experiment contained a total of 336 pairs, the second 360 pairs. Just noticeable differences were determined for each of the three contexts and each individual subject separately.

The fraction of stimuli reported different for any variable increment were adjusted for guessing. We assumed that the probability of a guess that the stimuli are different when in fact they are the same, is given by  $p(d/\Delta=0)$ , the fraction of different responses for the "same" stimulus pair. Following Swets (1964), the apparent probability of discrimination at some variable-difference value  $\Delta$  given by

$$p(d/\Delta) = p'(d/\Delta) + p'(d/\Delta=0) [1 - p'(d/\Delta)]$$

where  $p'(d/\Delta)$  is the true or corrected discrimination probability and  $p(d/\Delta=0)$  is the guessing probability. After the data were corrected for guessing, sigmoid curves were fitted to the data (Finney, 1964), and the points of 50 percent discriminability and their standard deviations were computed.

## Results

The DL values for Experiment 1 are given in Table 2. The direction above or below the center formant frequency in which the first or second formant was moved did not appear to affect the results significantly; therefore, the data are pooled and tabulated according to the formant varied. The mean DL for  $F_1$  variations is 50 Hz for the steady-state vowels and 49 Hz for the consonantal context. For none of the subjects is the difference in  $F_1$  discriminability due to context significant. The mean DL for  $F_2$  variations is 142 Hz for the steady-state vowel, 174, and 199 Hz respectively for the /b/ and /g/ contexts. The average increase in DL value is 45 Hz, or 31 percent over the DL value for the steady-state vowel. All the DL values in consonantal context are higher than the corresponding values for the same subject for the steady-state vowel. Six of the ten differences are significant at the .01 level. Significant variations are noted among the data of the various listeners. Note particularly the low DL values for subject 2 and the high values for subject 4 in all contexts.

TABLE 2: Difference limen and standard deviation values (Hz) for listeners to stimuli of Experiment I.

|                            | Subject  |          |          |          |          | Mean |
|----------------------------|----------|----------|----------|----------|----------|------|
|                            | #1       | #2       | #3       | #4       | #5       |      |
| Variation in $F_1$         |          |          |          |          |          |      |
| Vowel alone                | 74 ± 8   | 36 ± 5   | 47 ± 6   | 56 ± 6   | 39 ± 6   | 50   |
| /b/ context                | 61 ± 8   | 40 ± 5   | 41 ± 6   | 57 ± 7   | 29 ± 4   | 46   |
| /g/ context                | 64 ± 6   | 35 ± 2   | 46 ± 7   | 56 ± 6   | 53 ± 5   | 51   |
| Variation in $F_2$         |          |          |          |          |          |      |
| Vowel alone                | 166 ± 20 | 79 ± 15  | 174 ± 34 | 164 ± 19 | 125 ± 14 | 142  |
| /b/ context                | 207 ± 46 | 136 ± 15 | 222 ± 52 | 181 ± 12 | 126 ± 13 | 174  |
| /g/ context                | 245 ± 69 | 109 ± 14 | 208 ± 38 | 230 ± 73 | 206 ± 49 | 199  |
| $\Delta F_2 = 2\Delta F_1$ |          |          |          |          |          |      |
| Variation in $F_2$         |          |          |          |          |          |      |
| Vowel alone                | 66 ± 11  | 63 ± 11  | 54 ± 11  | 76 ± 10  | 63 ± 14  | 64   |
| predicted                  | 111      | 53       | 83       | 93       | 66       | 81   |
| /b/ context                | 142 ± 28 | 66 ± 11  | 80 ± 14  | 88 ± 12  | 63 ± 14  | 88   |
| predicted                  | 105      | 69       | 77       | 96       | 53       | 80   |
| /g/ context                | 101 ± 21 | 53 ± 11  | 80 ± 16  | 78 ± 19  | 63 ± 14  | 75   |
| predicted                  | 114      | 59       | 80       | 101      | 86       | 88   |
| Short vowel                |          |          |          |          |          |      |
| $F_1$ variation            | 46 ± 8   | 30 ± 7   | 54 ± 10  | 38 ± 5   | 42 ± 10  | 42   |
| $F_2$ variation            | 113 ± 22 | 63 ± 14  | 120 ± 30 | 134 ± 25 | 71 ± 21  | 100  |



To explore the effects of simultaneous variation in both the first and the second formant frequencies, variable stimuli were included in the DL tests that were constrained to lie along the line in formant-space given by  $\Delta F_2 = 2\Delta F_1$ ,  $\Delta F_2 = 25, 50, 75$  and  $100$  Hz, respectively. The DL values for  $F_2$  under these conditions are also given in Table 2 for each context. Let us assume a simple model that considers  $F_1$  and  $F_2$  to be independent parameters that contribute information to discriminability given by

$$\sqrt{w_1(\Delta F_1)^2 + w_2(\Delta F_2)^2}$$

where  $w_1$  and  $w_2$  are appropriate weighting factors. This model results in elliptical DL regions in formant space. On the basis of DL measurements for  $F_1$  and  $F_2$  variation alone, this model allows prediction of the DL when both parameters are varied simultaneously. The predicted DL values for  $F_2$  under the constraint that  $\Delta F_2 = 2 \Delta F_1$  are also given in Table 2. The measured DL values appear to be reasonably well predicted from the DL values for independent formant variation when  $F_1$  and  $F_2$  are not located close to each other.

One hypothesis that may account for the increased DL for the consonant-vowels embedded is that the time integral of the difference between the formant trajectories of the time-varying stimuli differing in a given  $F_2$  value, is less than the same difference for the steady-state stimuli. One may argue that if the temporal duration of a given difference is reduced, the discriminability may also be reduced. Liang and Chistovich (1961) found no substantial difference in the DL value for steady tones near 1 kHz between duration values of 100 and 200 msec. The duration had to be reduced below 100 msec in order for an increase in DL to be manifested. To verify the duration effects for speechlike stimuli, the DL was also measured for 133 msec long steady-state vowels around the same point in formant-space. No increase in DL value due to the shorter stimulus duration was noted for any subject. In fact, a decrease was measured in the second-formant DL value relative to the longer vowels that is significant at the  $p < .01$  level for three of five subjects. We know no good explanation for the causes of this reduction in DL values accompanying a reduction in stimulus duration. We may conclude, however, that the decreased discriminability is not due to the decreased average duration of the formant differences.

The results of the second experiment are shown in Table 3. Here differences were apparent in the DL values determined for variations in the positive and negative directions of the formant frequencies. Therefore, the data for these directions are shown separately. For each of the four directions of variations and for all contexts, the DLs in consonantal context are larger than the DLs for the steady-state vowels. All but one of the 48 differences are significant at the .01 level. The average DL for  $F_1$  is 33 Hz for steady-state vowels, 70 Hz in consonantal context. The average DL for  $F_2$  is 75 Hz for steady-state vowels, and 171 Hz in consonantal context. The average increase in each case exceeds 100 percent. The mean DL data for each vowel and context are summarized in Figure 2.

Figure 2: Just noticeable difference regions for vowels in and out of consonantal context. The results are mean values for five subjects (/i/ group) and six subjects (/ε/ group) respectively.

FIGURE 2

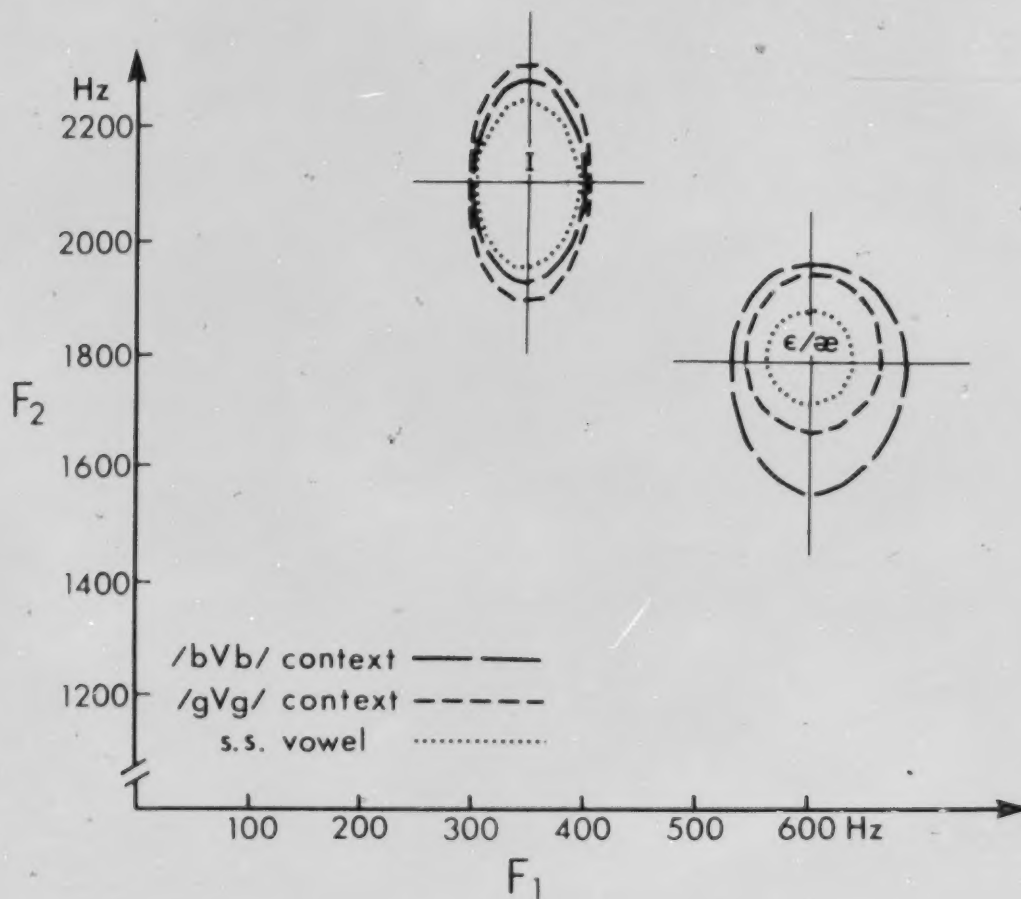


TABLE 3: Difference limen and standard deviation values (Hz) for listeners to stimuli of Experiment II.

|                                 | Subject  |           |          |          |          |          | Mean |
|---------------------------------|----------|-----------|----------|----------|----------|----------|------|
|                                 | #1       | #2        | #3       | #4       | #5       | #6       |      |
| <b>+F<sub>1</sub> variation</b> |          |           |          |          |          |          |      |
| s-s vowel                       | 37 ± 6   | 29 ± 7    | 32 ± 4   | 33 ± 5   | 39 ± 5   | 29 ± 4   | 33   |
| /b/ context                     | 67 ± 7   | 174 ± 112 | 41 ± 5   | 75 ± 6   | 83 ± 15  | 52 ± 6   | 82   |
| /g/ context                     | 83 ± 7   | 66 ± 9    | 38 ± 7   | 81 ± 6   | 72 ± 13  | 56 ± 8   | 66   |
| <b>-F<sub>1</sub> variation</b> |          |           |          |          |          |          |      |
| s-s vowel                       | 30 ± 5   | 40 ± 8    | 29 ± 5   | 42 ± 6   | 37 ± 4   | 27 ± 4   | 34   |
| /b/ context                     | 55 ± 5   | 82 ± 10   | 39 ± 7   | 77 ± 8   | 96 ± 16  | 85 ± 13  | 72   |
| /g/ context                     | 41 ± 6   | 87 ± 17   | 31 ± 5   | 88 ± 11  | 56 ± 8   | 60 ± 8   | 61   |
| <b>+F<sub>2</sub> variation</b> |          |           |          |          |          |          |      |
| s-s vowel                       | 62 ± 12  | 115 ± 18  | 60 ± 9   | 136 ± 19 | 107 ± 17 | 70 ± 12  | 92   |
| /b/ context                     | 182 ± 19 | 205 ± 42  | 114 ± 13 | 182 ± 19 | 189 ± 38 | 202 ± 36 | 179  |
| /g/ context                     | 161 ± 50 | 182 ± 35  | 113 ± 18 | 149 ± 17 | 186 ± 37 | 156 ± 16 | 158  |
| <b>-F<sub>2</sub> variation</b> |          |           |          |          |          |          |      |
| s-s vowel                       | 58 ± 10  | 35 ± 7    | 35 ± 7   | 93 ± 14  | 59 ± 12  | 68 ± 11  | 58   |
| /b/ context                     | 245 ± 80 | 269 ± 98  | 125 ± 14 | 233 ± 52 | 242 ± 72 | 261 ± 82 | 229  |
| /g/ context                     | 180 ± 20 | 105 ± 15  | 65 ± 10  | 151 ± 14 | 130 ± 19 | 92 ± 10  | 120  |

One additional experiment was carried out to test the comparative discriminability of vowels in isolation and in CVC context over larger region of the vowel space. Stimuli were regularly spaced along a line in the space of the first two formant frequencies such that F<sub>1</sub> was constant at 600 Hz, and F<sub>2</sub> ranged from 1 kHz to 2 kHz in 50 Hz increments. This allowed generation of vowels varying through /ɔ, ʌ, ø, ɛ, and æ/. The consonant in Experiment III was always /b/, achieved by using initial and terminal second formant frequency values ranging from 500 to 1000 Hz in 25 Hz steps. The initial and terminal F<sub>1</sub> frequency was always 50 Hz. All formant trajectories were parabolic in shape and had a total duration of 200 msec. The corresponding steady-vowels had the same duration.

In Figure 3 we have plotted the average discrimination by four subjects of pairs of stimuli spaced 3 steps, or 150 Hz in the central F<sub>2</sub> value. The short bars indicate the one standard deviation limit in each case. The overall average discrimination is 83 ± 14 percent for the vowel alone, but only 26 ± 22 percent for the vowel in the CVC context. The loss in

Figure 3: Three-step discrimination functions for vowels alone and in CVC context.

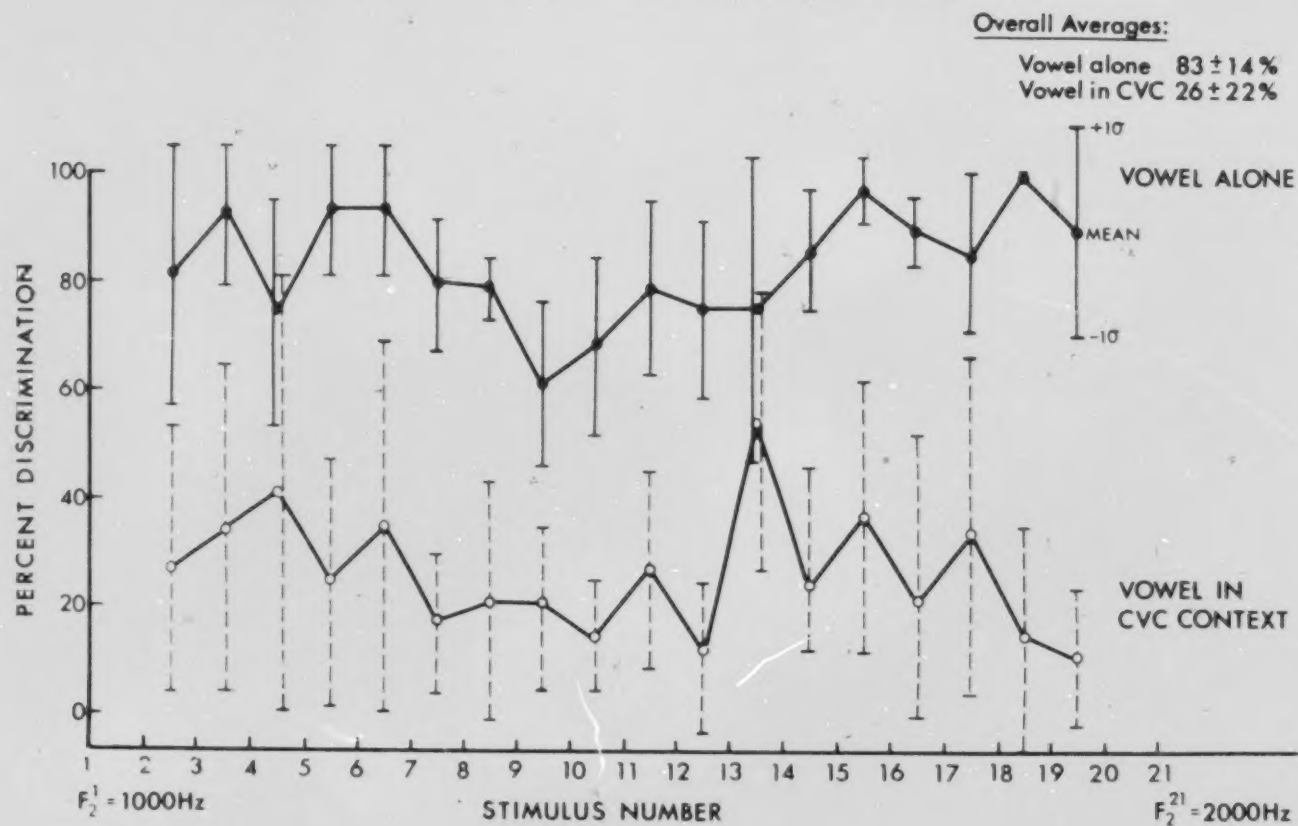
Three-step Discrimination Functions ( $\Delta F_2 = 150\text{Hz}$ )

FIGURE 3



discrimination due to the consonantal context is clearly demonstrated.

### Discussion

Comparison between Experiments I and II reveal a decrease in the steady-state vowel DL for  $F_1$  as the frequency of  $F_1$  increased. Similarly, there was a decrease in the steady-state vowel DL for  $F_2$  with the decreasing frequency of  $F_2$ . These results correlate well with the decrease in DL found by Flanagan (1955) when the formant amplitude increases due to the increased proximity of the two formants.

The DL value for  $F_1$  for the steady-state /i/ vowel appears significantly higher than the value reported by Flanagan (1955) for nearly the same  $F_1$  frequency. We obtained a value of 50 Hz; his result at an  $F_1$  of 300 Hz is roughly 15 Hz. At an  $F_1$  of 600 Hz, (steady-state /ε/ or /æ/, the difference is much smaller, 33 Hz vs. roughly 25 Hz at  $F_1$  values of 500 and 700 Hz by Flanagan. At formant-frequency values of 300, 500, and 700 Hz we have nearly equal excitation of two harmonics of the 120 Hz fundamental frequency. At 350 Hz and 600 Hz we have clear indication of a maximally excited harmonic. If discrimination requires a shift in the particular harmonic that is maximally excited, then DL values about a formant-frequency value that straddles two harmonics should be smaller than DL values about a formant-frequency value that corresponds more closely to a specific harmonic. The location of the formant-frequency relative to the harmonics may therefore account for the cited differences in measured DL values. In addition, there may be a slight increase in the DL value due to the use of a final drop in the fundamental frequency contour in our experiments, as compared to a steady fundamental in case of Flanagan's study.

When we compare the set of DL values measured in the two experiments, we find that the values in the consonantal context are larger for both  $F_1$  and  $F_2$  in Experiment II. The extent of formant transitions follows the same relative order. The lack of a significant difference in the DL for  $F_1$  in Experiment I in and out of context may be due to the small range of  $F_1$  transition there.

We noted a further trend toward dependence of the DL value for  $F_2$  on the particular consonant. In 6 of 12 cases (6 subjects, 2 directions), the difference due to /g/ or /b/ context was significant. The sign of this difference was reversed for positive and negative variations in  $F_2$  in each case.  $F_2$  variations in /b/ context are concave downward. For this case, the DL for positive  $F_2$  variation is smaller than the DL for negative  $F_2$  variation. However,  $F_2$  variations in /g/ context are concave upward. Here the DL for positive  $F_2$  variation is larger than the DL for negative  $F_2$  variation. In each case the larger DL value is in the direction of the terminal  $F_2$  value.

The above difference in discriminability may be accounted for by a boundary proximity effect. If we postulate an /ε/-/æ/ phoneme boundary above the central  $F_2$  frequency for the /g/ context, and below the central  $F_2$  frequency for the /b/ context, the boundary shift would be in a direction consistent with the findings of Lindblom and Studdert-Kennedy (1967). In other words, there may be a boundary shift due to perceptual compensation for expected undershoot in vowel articulation. Moving  $F_2$  in the direction of

that boundary may result in a lower DL, than moving it away from the boundary may.

To test this hypothesis, a vowel-identification experiment was run with five additional subjects on the stimuli used in the second experiment. The subjects used in the original discrimination experiment were no longer available. The stimuli varying in the central  $F_2$  value for steady-state, /b/ and /g/ contexts were used. These stimuli could be identified as /ε/, /æ/, or "other vowel." Within the limits of the variation among the stimuli, no crossover was observed for any subject between a majority of /æ/ and a majority of /ε/ observations along any of the stimulus continua. The asymmetry in /ε/-/æ/ identification can be expressed as

$$(N_{\epsilon}^{+} - N_{\epsilon}^{-}) - (N_{\epsilon}^{-} - N_{\epsilon}^{+}) / (N_{\epsilon}^{+} + N_{\epsilon}^{-} + N_{\epsilon}^{+} + N_{\epsilon}^{-})$$

where  $N_{\epsilon}^{+}$  is the number of positive  $F_2$  excursion stimuli identified as /æ/,  $N_{\epsilon}^{+}$  the number of positive  $F_2$  excursion stimuli identified as /ε/, and  $N_{\epsilon}^{-}$  and  $N_{\epsilon}^{-}$  are defined similarly for the negative  $F_2$  excursion stimuli. Existence of a perfect boundary at the formant values of the standard would have resulted in a ratio of 1. For no subject did this ratio exceed 0.20. Therefore it is unlikely that, for a fixed stimulus duration of 200 msec, the 400 Hz range of  $F_2$  variation was sufficient to significantly effect the probability of /ε/ vs. /æ/ judgment. The difference in discriminability between the positive and negative directions of formant variation that we have found appears not to be due to proximity of a vowel boundary.

The results of the third experiment indicate that discrimination of the steady-state vowel series is clearly continuous. There appear two minor peaks in the CVC discrimination curve at 1175 and 1625 Hz, and these roughly correspond to the /ɔ/-/Δ/ and /Δ/-/ε/ boundaries as observed from the identification of the same stimuli. However, discrimination is not sharply categorical, and the identification curves of the different vowels show significant overlap. Furthermore, discrimination does not vary appreciably with the central  $F_2$  value, as that value is changed from 1000 to 2000 Hz. Thus, it is unlikely that the critical bandwidth of the auditory system, that increases as we traverse this frequency range, plays a significant role in establishing the difference-limen value.

### Conclusions

Formant-frequency differences that are discriminated among steady-state vowels are not discriminated in the presence of consonantal transitions. We interpret this result as due to a stage in processing the incoming speech signal where some information concerning formant differences is lost. Two mechanisms may account for this result.

First, we may consider the result as purely auditory in origin. Steady-state vowels are mapped into a unique excitation pattern on the basilar membrane. Any time variation of the vowel spectrum can be expected to broaden the excitation pattern and thus add some uncertainty to the spectral component of perceived vowel quality.

Tsumura, Sone and Nimura (1973) carried out experiments on the just noticeable differences between two steady portions of a tone burst separated by rising and falling transitions. The threshold frequency difference increased roughly logarithmically, with decreasing transition duration for transitions shorter than 50 msec. For tones near 1 kHz in frequency, the threshold was found to be greater when the frequency transition occurs near the burst onset, than if it occurs near the cutoff. The threshold was generally higher for a rising transition than for a falling transition. Evidently there is an auditory masking effect of a steady tone in the presence of frequency transitions. The frequency transitions in our data are significantly larger than those used in the Tsumura study, thus masking effects of significantly greater magnitude may occur in speech than those reported for tones. These effects would tend to increase the just noticeable difference (JND) of the formant frequencies in context above those that hold for steady-state vowels.

Second, the effect may be due to the phonetic decoding that must follow the auditory processing of the CVC stimuli. Liberman, Mattingly, and Turvey (1972) suggest that phonetic decoding of consonants "strips away auditory information" in short-term memory. A similar process may be involved in the decoding of vowels as well, but perhaps information is lost at a slower rate. Discrimination in consonantal context is still quite good despite the fact that the vowels belong to the same phoneme category. Yet it is not as good as that observed for the steady-state vowels. A partial loss of information has taken place.

The improved identification for natural vowels in context over those in isolation obtained by Strange, Verbrugge, Shankweiler, and Edman (1976) are apparently not due to improved formant frequency discrimination. Our data imply that on the basis of frequency information alone, steady-state vowels are better discriminated than context-embedded vowels. However, if temporal information concerning the differences in the transitions is used in the discrimination task in addition to the formant-frequency differences, the consonant-embedded vowels could be better discriminated. Durational information, in particular, was missing from the /ε/-/æ/ stimulus series. This durational information that is regularly used when interpreting continuous speech (Peterson and Lehiste, 1960; Klatt, 1976), may more than suffice to overcome the reduced frequency discrimination, and thereby result in improved identification of vowels in consonantal context. The results suggest that improved vowel-recognition performance in automatic speech recognition is not to be attained by improved accuracies in formant frequency determination. Rather, contextual and temporal factors must be utilized as well.

At an  $F_1$  value of 600 Hz, the 1K-2K frequency range for  $F_2$  covers four different English vowels, /ɔ/, /ʌ/, /æ/, and /ε/. We have measured a DL for  $F_2$  in consonantal context of nearly 200 Hz. One may argue that in view of the large DL value, no more than five vowels can be reliably differentiated in this frequency range. Liljencrants and Lindblom (1972) discuss a 12-vowel model where the vowels are distributed in  $F_1$ - $F_2$  space on the basis of maximum perceptual contrast. Even for that system, no more than four vowels are located in the above frequency range. It appears that there exist fundamental auditory and phonetic limitations on the density of vowels in formant space, and vowel-rich languages such as English have closely approached those limits.

The results concerning increased DL values in consonantal context appear to have significant implications for speech coding applications. The limited data indicate an increase that frequently exceeds 100 percent for the DL values in consonantal context. More detailed exploration of the limitations of these results, if any, needs to be carried out. In particular, one must determine whether the same results are obtained throughout the vowel space and in all consonantal contexts. If the results can be generalized to all vowels and contexts, then when encoding syllable-sized units, we need to quantize the formant-frequencies of the syllabic peak (the point of maximal spectral stability) with a resolution that may only be half as fine as criteria based on the DL values for steady-state vowels. Since spectral discrimination at points of greater spectral variation can be expected to be worse, discrimination at the syllabic peak appears to pose the tightest requirements.

The DL values cited cannot be applied directly for the independent encoding of formant information for successive short time segments. Clearly, independent perturbations of formant-frequency values of a magnitude comparable to the measured DL values would be unacceptable. However, where syllable-sized segments are encoded as a unit, in terms of duration and spectral parameters, substantial information-rate savings may be achievable.

#### REFERENCES

- Finney, D. J. (1964) Probit Analysis - A Statistical Treatment of the Sigmoid Response Curve. (Cambridge: Cambridge University Press).
- Flanagan, J. L. (1955) Difference limen for vowel formant frequency. J. Acoust. Soc. Am. 27, 613-617.
- Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Lang. Speech 5, 171-189.
- Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute, (University of Tokyo) 29, 207-214.
- Klatt, D. H. (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. J. Acoust. Soc. Am. 59, 1208-1221.
- Liang, L. C. and L. A. Chistovich. (1971) Frequency difference limens as a function of tonal duration. Soviet Physics-Acoustics 6, 75-80.
- Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D.C.: V. H. Winston).
- Liljencrants, J. and B. Lindblom. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. Language 48, 839-862.
- Lindblom, D. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. 42, 830-843.
- Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of vowels. J. Acoust. Soc. 24, 175-184.
- Peterson, G. E. and I. Lehiste. (1960) Duration of syllable nuclei in English. J. Acoust. Soc. Am. 32, 693-703.
- Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. (Ph.D. thesis, University of Michigan). [Supplement to Haskins



Laboratories Status Report on Speech Research].

- Rabiner, L. R. (1968) Digital-formant synthesizer for speech-synthesis studies. J. Acoust. Soc. Am. 43, 822-828.
- Stevens, K. N. (1968) On the relations between speech movements and speech perception. Z. Phon. Sprachwiss. u. Komm. Forschung 21, 102-106.
- Stevens, K. N. and A. S. House. (1963) Perturbations of vowel articulations by consonantal context: An acoustical study. J. Speech Hear. Res. 6, 111-128.
- Stevens, K. N., A. S. House, and A. P. Paul. (1966) Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. J. Acoust. Soc. Am. 40, 123-132.
- Strange, W., R. R. Verbrugge, D. P. Shankweiler, and T. R. Edman. (1976) Consonant environment specifies vowel identity. J. Acoust. Soc. Am. 60, 213-224.
- Swets, J. A. (1964) Signal Detection and Recognition by Human Observers. (New York: J. Wiley).
- Tsumura, T., T. Sone, and T. Nimura. (1973) Auditory detection of frequency transition. J. Acoust. Soc. Am. 53, 17-25.

ABSTRACT

An acoustic cue separating /s/ and /š/ was studied with synthetic speech stimuli, in the context of two different following [æ]'s: one appropriate for a large vocal tract and one for a small vocal tract. For each vocal tract size, the center frequency of the frictional portion of the stimuli was varied in ten steps from 2974 to 4455 Hz. In preliminary experiments, it was found that listeners have a reasonably consistent boundary along this scale at which they stop hearing /š/ and start hearing /s/. It was hypothesized that a listener's boundary would be higher for the small-vocal-tract stimuli than for the large-vocal-tract stimuli. Nineteen subjects, who were asked to identify 40 tokens of each stimulus, showed an upward boundary shift of varying magnitude for the small-vocal-tract stimuli as compared with the large-vocal-tract stimuli. This result suggests the /s/-/š/ cue is dependent on vocal tract size.

For many years, one of the most intriguing characteristics of the speech signal has been its lack of acoustic invariance, resulting from differences in vocal tract size and shape, dialect, rate of articulation, and stress. This was first suggested by the data of Peterson and Barney (1952), who showed that tokens of the same American English vowel spoken by men, women, and children may yield different acoustic specifications. Subsequently, Ladefoged and Broadbent (1957) found that listener's identifications of certain synthetic vowels shifted with the vowel space of the context in which they were embedded. In 1971, Rand observed a similar phenomenon in voiced stops: he showed that the formant transition boundary values between synthetic /b/ and /d/, /d/ and /g/ shifted upward when heard in the context of a vowel from a smaller vocal tract. The present study seeks to demonstrate a comparable lack of acoustic invariance for the fricatives /s/ and /š/.

The fricatives /s/ and /š/ are characterized by a noisy portion centering around a particular frequency and by formant transitions to preceding and following vowels. In the work presented here, however, only the cue provided by the noise is considered. Since the noisy portion of /s/ and /š/ are more prominent perceptual cues than their transitions (Harris, 1958), it is possible to create acceptable synthetic /s/ and /š/ with noisy

---

\*University of Connecticut, Storrs.

portions, but without transitions.

Hughes and Halle have shown that, in general, the band of fricative noise for /s/ starts above 3500 Hz and extends to over 8000 Hz, while for /š/ the band starts between 1600 and 2500 Hz and rarely exceeds 7000 Hz. Spectrographic analysis of /s/ and /š/ spoken by one male and one female suggested that the noise frequencies of both phonemes vary with vocal tract size. On the average, the fricative noise for /š/ began at 1800 Hz for the male compared to 2600 Hz for the female, while for /s/ it began at 3700 Hz for the male and at 5100 Hz for the female.

Preliminary experiments revealed that when listeners are presented with randomized synthetic fricatives, cued by noise bands of about 200 Hz in width with center frequencies varying from 2974 Hz to 4455 Hz, they exhibit reasonably consistent boundaries, somewhere between these two frequencies, at which they stop hearing /š/ and start hearing /s/. It was hypothesized that the noise boundary for a fricative followed by a synthetic vowel would be higher for a vowel apparently produced by a small vocal tract than for a vowel apparently produced by a large vocal tract.

#### METHOD

##### Stimuli

In the experiment, there were two different groups of synthetic consonant-vowel stimuli, one representative of a large vocal tract, and one of a small vocal tract. Figure 1 displays schematic spectrograms of the stimuli. Each stimulus consisted of a period of friction typical of /s/ or /š/ followed by, in the group on the left, a steady-state /æ/ representative of a large male vocal tract, and in the group on the right, an /æ/ representative of a small male vocal tract. There were no formant frequency transitions.

All vowels lasted 300 msec and had fundamental frequencies of 156 Hz. The only difference between the two groups of stimuli was in the formant frequencies of the vowels. The formants for the small vocal tract /æ/ were taken from the average values for female /æ/ reported by Peterson and Barney (1952): a first formant of 850 Hz, a second formant of 2035 Hz, and a third formant of 2813 Hz. However, the resulting synthetic /æ/ sounded less like a female /æ/ than like an /æ/ produced by a small male vocal tract. The formants for the large vocal tract /æ/ were related to those of the small vocal tract by a multiplicative constant of .8 which yielded a first formant of 673 Hz, a second formant of 1641 Hz, and a third formant of 2410 Hz; these values are close to the average values for male /æ/ reported by Peterson and Barney (1952).

For both groups of stimuli, the center frequency of the friction was varied in ten equal steps from 2974 to 4455 Hz. The lower dotted areas in Figure 1 represent the stimuli with noisy portions centered at 2974 Hz, while the upper ones represent stimuli with noisy portions centered at 4455 Hz. Each friction portion was 200 msec long, had a bandwidth of about 200 Hz, and an amplitude of -16.5 dB relative to the amplitude of the vowel's first

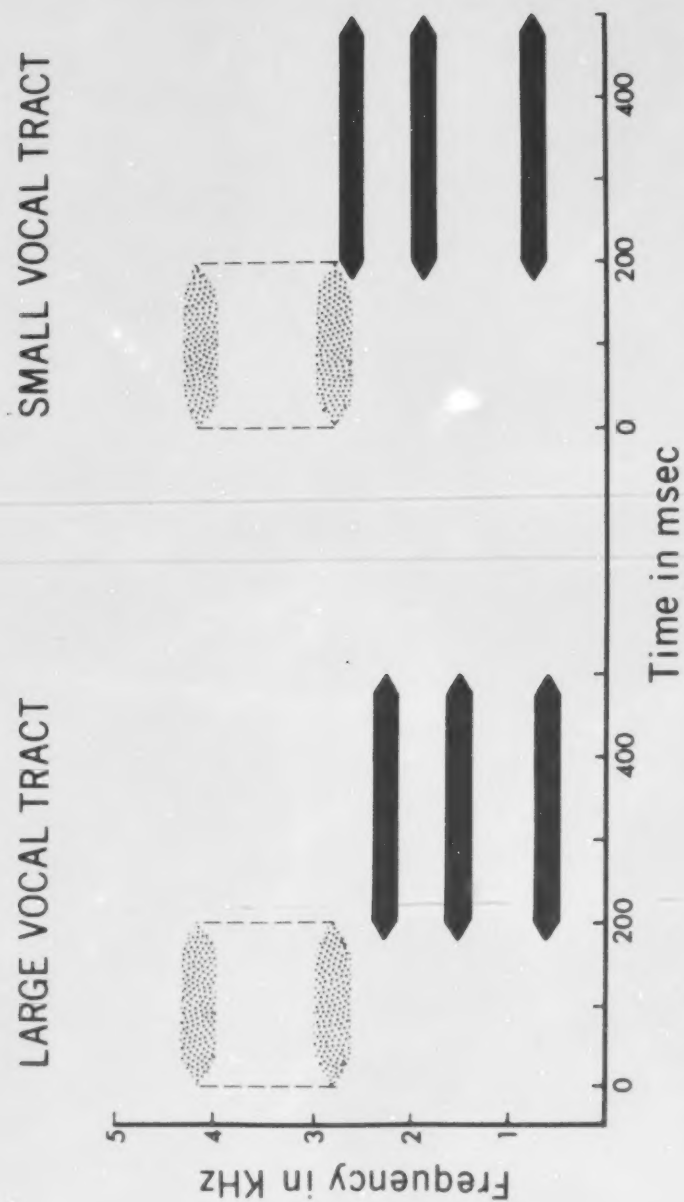


Figure 1: Schematic spectrograms of a synthetic fricative continuum, ranging from /s/ to /š/ and followed by the vowel /a/, synthesized with formant frequencies appropriate to either a large or a small vocal tract.



formant. All stimuli were made by a Glace-Holmes terminal analog synthesizer.

### Experimental Tests

From each group of ten stimuli, two randomizations were made. Each randomization contained 100 consonant-vowel syllables, ten of each of the ten stimuli in that group. They were arranged in blocks of ten stimuli eight seconds apart. Within each block the interstimulus time was two seconds.

Two pretests were also prepared--one for each vocal tract size--in order to acquaint the subjects with synthetic stimuli.

### Subjects

Nineteen graduate and undergraduate students from the University of Connecticut were used as subjects. They were phonetically naive, and to the best of their knowledge had no hearing difficulties. All were native speakers of English. Eight were female, eleven male.

### Procedure

A subject's task was to listen to each stimulus and identify it as either /sæ/ or /ʃæ/. They heard the two pretests first, followed by the four randomizations described above. This took about one-half hour. After a three to five minute break, subjects heard the same four tests again. Thus the number of responses from each subject for each of the 20 stimuli was 40.

## RESULTS

### Labeling Results

Figure 2 presents the pooled labeling functions for all nineteen subjects for the small vocal tract condition on top, and for the large vocal tract condition on the bottom. The group /s/-/ʃ/ boundary falls between the sixth and seventh stimuli for the small vocal tract data, between the fourth and fifth stimuli for the large vocal tract data.

The identification curves for most subjects represented in these figures had steep slopes, indicating sharply defined consonant categories. Only one subject (JL, who also showed the largest boundary shift) gave labeling functions with gradual slopes: in neither the small nor large vocal tract tests did he identify the extremes of fricative noise as /s/ or /ʃ/ 100 percent of the time. However, since the exact position of the boundary varied from subject to subject (see Figure 3), the group labeling functions are not very steep.

### Boundary Shifts

The category boundary was defined as the point along the continuum of noise center frequencies where 50 percent of the time a stimulus is heard as

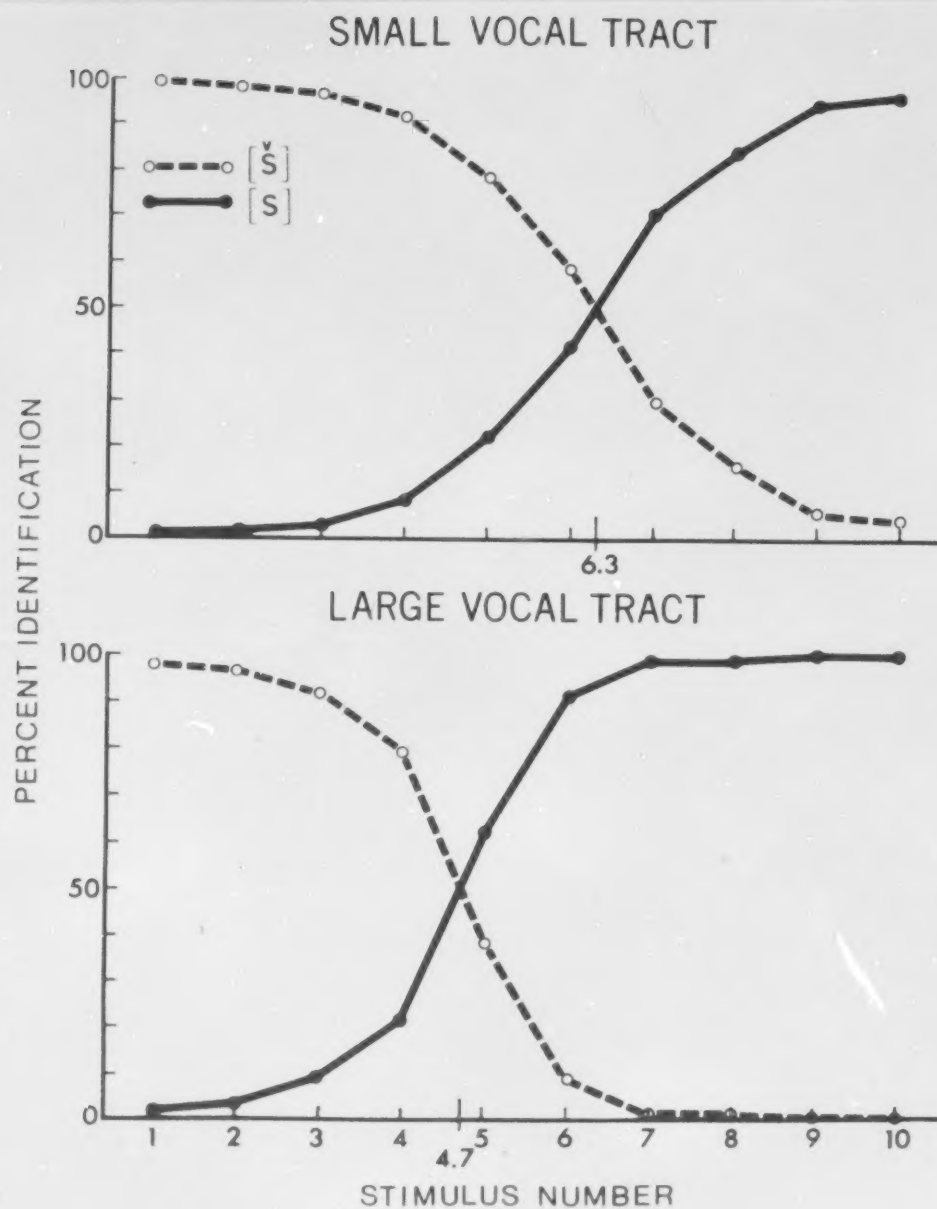


Figure 2: Pooled labeling functions for nineteen subjects.

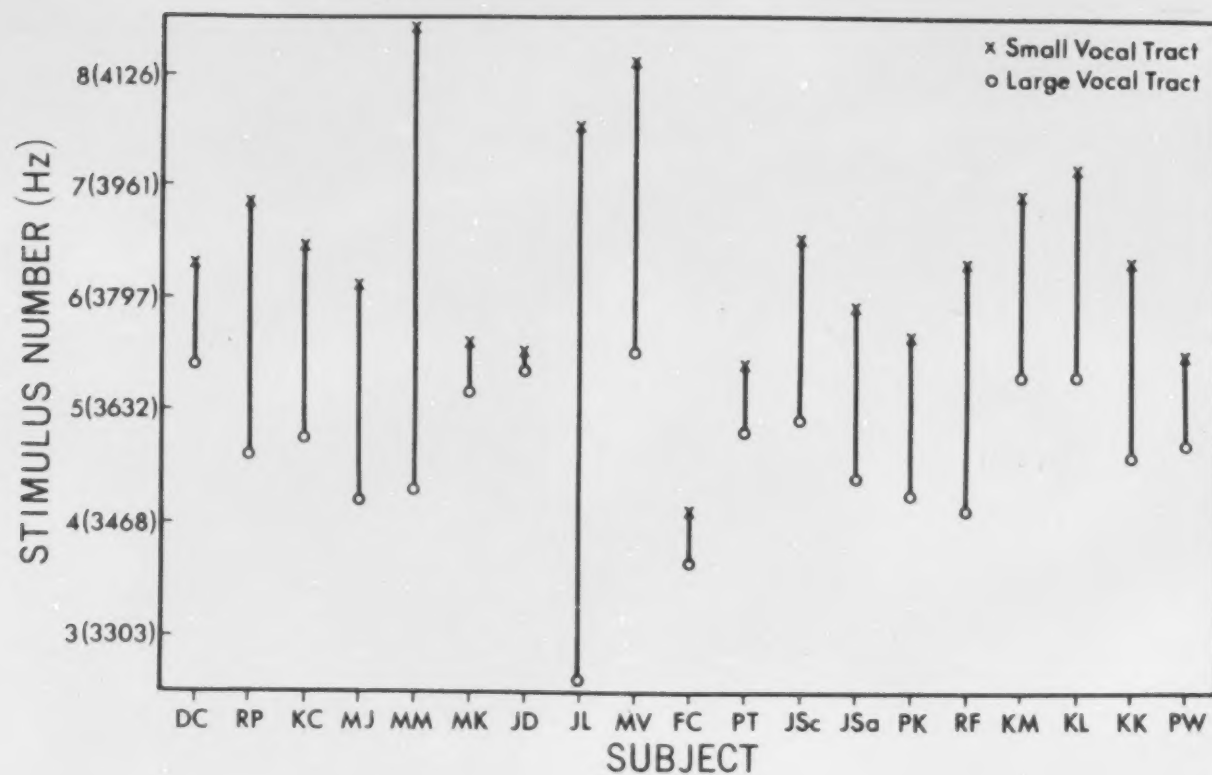


Figure 3: Boundary shifts for all subjects.

/s/, 50 percent as /š/, or where the /s/ and /š/ curves intersect in a figure such as Figure 2. The boundary shifts for individual subjects are shown in Figure 3. All nineteen subjects showed boundary shifts in the same direction--from low along the friction continuum for the large vocal tract (represented by the circles) to high for the small vocal tract (represented by the crosses). There were two subjects--MK and JD--who showed shifts within a stimulus step. There were four subjects, DC, FC, PT, and PW, who showed boundary shifts of a magnitude smaller than a full stimulus step. The magnitude of the shifts of all subjects ranged from less than one stimulus step for MK and JD, to five stimuli steps for JL, with the majority of the shifts having a magnitude of one to two stimuli steps.

#### CONCLUSION

Probable explanations for the varying sizes of subjects' boundary shifts include individual differences in task strategy and in ability to accept synthetic stimuli, combined with the fact that, due to the design of the experiment, the only available normalization cue was in the following vowel. Presumably natural speech would carry many more cues.

Nevertheless, these data provide convincing evidence of a boundary shift. They indicate that /s/ and /š/ do not have invariant acoustic cues, and that the cues vary as a function of vocal tract size. In this experiment information carried by the following vowel seems to have determined the position of the acoustic boundary between these fricatives. One may conclude that vocal tract normalization occurs for these fricative consonants as well as for the vowels and stop consonants previously studied.

#### REFERENCES

- Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. Lang. Speech, 1, 1-7.
- Hughes, G. W. and M. Halle. (1956) Spectral properties of fricative consonants. J. Acoust. Soc. Am. 28, 303-310.
- Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Am. 29, 98-104.
- Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of vowels. J. Acoust. Soc. Am. 5, 175-184.
- Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research SR-25/26, 141-146.



ABSTRACT

In our studies of reading acquisition, we have been guided by the assumption that reading is not an independent ability, but is dependent upon speech. We have explored the implications of the dependence of reading on speech in three related groups of investigations. First, since an alphabetic writing system is a more or less phonetic representation of the spoken language, it seemed obvious that in order to map the written word to the spoken word, the child must have some recognition of the phonetic structure of his spoken language. There is reason to suspect that the development of the awareness of phonemic segments might be difficult for the child. Accordingly, we studied the child's development of phoneme segmentation and the relation of this ability to reading. Second, if reading is dependent upon speech, then it seems likely that reading relies on many of the same mental processes as speech perception. Since the role of the phonetic representation in speech perception is to hold information about shorter segments in short-term memory until the meaning of the longer segments can be extracted, we were led to wonder if the phonetic representation derived from optical information might not serve the same purpose in reading. We investigated the use of phonetic coding by beginning readers and found differences in this ability between good and poor readers. Third, we obtained a corpus of children's reading errors in order to test our hypothesis about linguistic factors in beginning reading. It was found that errors on consonant and vowel segments pattern differently. We believe that this may reflect the

---

\*This paper was presented at the Conference on the Theory and Practice of Early Reading Instruction, Learning Research and Development Center, University of Pittsburgh, 21 May, 1976, and will be published in the Conference Proceedings: Theory and Practice of Early Reading, ed. by L. Resnick and P. Weaver. (Hillsdale, New Jersey: Lawrence Erlbaum Assoc.).

†University of Connecticut, Storrs, Connecticut.

††Also University of Connecticut, Storrs, Connecticut.

Acknowledgement: We are indebted to our colleague, Robert Verbrugge, for a critical reading of the manuscript.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

different linguistic functions of consonants and vowels. At all events, the error pattern cannot be attributed to the optical characteristics of the letters that represent these phonetic categories. Last, we suggest some ways in which our research findings can be applied to the teaching of reading; we also present some observations about the possible contribution of the orthographic complexity of English to the problems of learning to read.

### INTRODUCTION

In the research we have done on reading acquisition, we have been governed by the assumption that reading is somehow parasitic on speech (Liberman, 1971, 1973; Shankweiler and Liberman, 1972). We have been led to this assumption by certain observations that seem obvious. First, speech is unquestionably the primary language system, naturally and universally acquired without direct instruction. Reading, being secondarily derived from speech, is relatively unnatural, far from universally learned, and must be taught. Second, an alphabetic writing system is a more or less phonetic representation of the spoken language; it is not a separate symbol system that is keyed directly to meaning. And finally, speech appears to be an essential foundation for the acquisition of reading. Children who are blocked in the acquisition of speech, like the congenitally deaf, do not readily learn to read even though they have access to the printed word through the visual channel.

In an effort to explore the implications of the dependence of reading on speech, we have, in our studies of young children, investigated three related aspects of the problem: linguistic awareness of phoneme segmentation, phonetic coding in short-term memory, and the phonetic pattern of reading errors.

In order to learn to read, the child must map the written word to the spoken word. It has seemed plain to us that to do this, he must have some recognition of the phonetic structure of his spoken language (Liberman, 1971). We know from speech research that phonetic structure is complexly encoded in the speech signal (A. M. Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). The consequence is that there is no obvious acoustic criterion that marks the phonemic segments. We were thus led to ask whether the development of the awareness of these segments might be difficult for the child. Accordingly, we have investigated the child's development of phoneme segmentation ability and the relation of this ability to reading (Liberman, 1973; Liberman, Shankweiler, Fischer, and Carter, 1974).

The role of the phonetic representation in speech perception is to hold information about shorter segments (say, words) in short-term memory until the meaning of longer segments (say, sentences) can be extracted. That has led us to wonder if the phonetic representation derived from optical information might not serve the same purpose in reading. Therefore, we have investigated the use of phonetic coding in reading and, particularly, the differences in this ability between good and poor readers (I. Y. Liberman, Shankweiler, A. M. Liberman, Fowler, and Fischer, 1976; Shankweiler and Liberman, 1976).

Finally, guided by what we hope is common sense, we have supposed that by noting the particular errors that beginning readers make and analyzing them appropriately we might gain some insight into the processes underlying reading acquisition, and at the same time, test our hypotheses about linguistic factors in beginning reading (Fowler, Liberman, and Shankweiler, 1976).

In this paper we will review our findings in these three areas of investigation, emphasizing recent work, and will suggest some of the ways in which these findings can be applied in reading instruction. We will conclude with some observations about the possible contribution of the orthographic complexity of English to the problems of the beginning reader.

### LINGUISTIC AWARENESS AND THE ALPHABET

In languages that are written alphabetically, the unit characters--letters--are keyed to the phonological structure of speech. We are aware that the mapping from written symbols to phonemes is more nearly one-to-one in other alphabetic languages--such as Finnish and Serbo-Croatian--than in English. The many departures from one-to-one mapping makes English difficult to spell, and probably more difficult to learn to read, than is the case in languages whose alphabetic writing systems have a simpler structure. We defer matters concerning the role of the orthography in the acquisition of reading until later sections of this paper. For the present, it is sufficient to underscore the fact that English spelling, in common with other orthographies that employ an alphabet, is, for all its peculiarities, a cipher on the phonemes of the language.

The child's fundamental task in learning to read is to construct a link between the arbitrary signs of print and speech. We have pointed out (Liberman et al, 1976) that there are different ways in which the child might do this. Words written by an alphabet can be read as though they were logograms, and many children undoubtedly begin reading in this way, apprehending the word shapes holistically, rather than analyzing them as letter strings. However, the reader who employs a nonanalytic strategy of this sort cannot benefit from a unique advantage of alphabetic writing. We refer to the fact that the alphabet enables its users to generate a word's pronunciation from its spelling. Thus a user can recognize in print a word he has never before seen written down, and he can (at least to a rough approximation) pronounce a word that he has never before either heard or read. These powerful advantages are open only to a user who knows how the alphabet works, that is to say, one who can approach the reading task analytically.

Let us outline briefly what is involved in analytic reading. First, the child must realize that speech can be segmented into phonemes and he must know how many phonemes the words in his vocabulary contain and the order in which they occur. Second, he must know that the letter symbols represent phonemes, not syllables or some other unit of speech.

In our earlier writings (Liberman, 1971; Shankweiler, and Liberman, I., 1972, 1976; Liberman, et al., 1976) we have considered what it means for a child to know that speech can be segmented into phonemes. It does not mean simply that the child is able to discriminate word pairs that are minimally different. Every normal child of school age can do that. However, a child

may be able to discriminate between pairs of spoken words such as bet and best and to recognize each as a distinct word in his vocabulary, without being aware that bet contains three phonemes and best contains four. Such a child, as we have said elsewhere, has only a tacit awareness of phoneme segmentation. This is sufficient, of course, for comprehension of the spoken message. Writing and reading, on the other hand, demand an additional capacity to analyze words as strings of phonemes. Mattingly (1972) and others have called this capacity "linguistic awareness."

We have suggested that an understanding of the acoustic structure of speech can help to explain why the ability to analyze syllables as strings of phonemes is rather difficult to attain (Liberman, 1971). We suspect that the difficulty has, in part, to do with the fact that phonemes are not represented in the acoustic signal in discrete bundles, but rather are merged--"encoded"--into the structure of the syllable (as suggested by A. M. Liberman, et al., 1967). The word dig, for example, has three phonetic segments but only one acoustic segment.

This merging of phonemes in the sound stream complicates the process of becoming actively aware of the phonemic level of speech for the would-be reader. We do not mean to imply that the young child has difficulty differentiating word pairs, such as bad and bat, that differ in only one phoneme. On the contrary, there is reason to believe that most children hear these differences as accurately as adults (Read, 1971). As we have said, the problem is not to teach the child to discriminate minimally different word pairs, but rather to bring him to realize that each of these words contains three segments, and that they are alike in the first two and different in the third.

Elsewhere (Shankweiler and Liberman, I., 1976), we have dwelt on another important consequence of the encoded nature of phonemes that must contribute to the difficulty of learning to read analytically. Since the syllable, and not the phoneme, is the minimal unit of articulation, it is impossible to read by sounding out the letters one by one. On the contrary, it is necessary to discover how many of the letter segments must be taken simultaneously into account in order to arrive at the correct phonetic rendition of each syllable. Thus, we have stressed that to read analytically is not to read letter by letter, even in languages in which the letter-to-sound mapping is more direct than in English.

We have argued that effective use of an alphabet requires a degree of active awareness of phonological structure that goes beyond the tacit level of comprehension adequate for speaking and listening. As we have seen, it is one thing to understand and to speak one's language and quite another thing to have analytic understanding of the language's internal structure. We have noted (Liberman, 1973) that the late appearance of the alphabet in the history of writing may be an indication that it is rather difficult to become aware of the phonological underpinnings of speech. If the obscurity of phoneme segmentation is a psychological fact about speech that indeed is related to the late appearance of alphabets, then it is reasonable to suppose that the child might find phonemic segmentation difficult. There might be in the development of the child an order of difficulty of segmentation from word to syllable to phoneme that parallels the historical development of writing systems.



### Development of the Awareness of Speech Segments in the Young Child

We tested the supposition directly in a recent experiment. The point was to determine how well children in nursery school, kindergarten, and first grade (4-, 5-, and 6-year-olds) can identify the number of phonetic segments in spoken utterances and how this compares with their ability to deal similarly with syllables (Liberman, et al., 1974). The procedure was in the form of a game that required the child to indicate, by tapping a wooden dowel on the table, the number (from one to three) of the segments (phonemes in the case of one group, syllables in the other) in a list of test utterances.

At age four, none of the children could segment by phoneme, whereas nearly half could segment by syllable. Ability to carry out phoneme segmentation successfully did not appear until age five, and then it was demonstrated by less than a fifth of the children. In contrast, almost half of the children at that age could segment syllabically. Even at age six, only 70 percent succeeded in phoneme segmentation, while 90 percent were successful in the syllable task.

### Segmentation Ability and Reading Skill

The difficulty of phoneme segmentation has been remarked by a number of investigators besides ourselves (Rosner and Simon, 1971; Calfee, Chapman, and Venezky, 1972; Gleitman and Rozin, 1973; Savin, 1972; Elkonin, 1973; Gibson and Levin, 1975). Their observations, like ours, also imply a connection between the awareness of phoneme segmentation and early reading acquisition.

We explored this question in a preliminary way by measuring the reading achievement of the children who had taken part in our experiment on phoneme segmentation the year before. Testing our first graders at the beginning of their second school year, we found that half the children in the lowest third of the class in reading achievement had failed the segmentation task the previous June; on the other hand, there were no failures in phoneme segmentation among children who scored in the top third of the class in reading ability (Liberman et al., 1976). Rosner (1975) has also found that the partial phoneme segmentation required by his elision task is also a significant predictor of reading achievement.

Three new studies by our research group now confirm these results. Despite widely varying school populations and diverse procedures, each of these studies shows a high and significant correlation between phoneme segmentation and early reading ability.

Helfgott (1976) recently completed a study of the segmentation and blending skills of kindergarten children in a white, middle-class suburban school in Connecticut. In connection with this study, she looked at the usefulness of several different skills as predictors of first grade reading achievement in the following year. Using an adaptation of the Elkonin procedure (1973) for the assessment of phoneme segmentation, she found that the best predictor was the ability to segment spoken consonant-vowel-consonant (CVC) words into their three constituent phonemes. The correlation of this ability with reading achievement on the word recognition subtest of the Wide Range Achievement Test (WRAT) (Jastak, Bijou and Jastak, 1965) was substantial ( $r = .75$ ).

In an investigation of the phonological awareness and reading acquisition of first graders in an integrated city school in Rhode Island, Zifcak (1976) demonstrated a highly significant relationship between ability to segment phonemes on the Liberman tapping task (Liberman et al., 1974) and reading success as measured by the Gallistel-Ellis Test of Coding Skills (1974), as well as on the WRAT.

Treiman (1976) examined first and second graders in an inner-city school with a largely black population in New Haven. She used a task requiring the placement of the correct number of tokens (rather than dowel tapping) to indicate the number of the constituent phonemes. Her stimuli were not words, but two- and three-segment syllables in which the incidence of eight vowels, four stops, and four fricatives was carefully equated. She added the much needed control of ascertaining the counting ability of the subjects. In addition to the WRAT, she included an experimenter-devised reading test that allowed for a more analytic assessment of early reading skills. Once again, in spite of these many variations, the relationship between segmenting ability and reading success was highly significant.

These investigations of the relation between segmentation abilities and reading proficiency suggested that ability to analyze speech phonemically is indeed relevant to success or failure in learning to read. The results so far lend encouragement to our hypothesis that segmentation abilities are cognitive prerequisites for reading. We turn now to consider another aspect of language development that may bear on reading acquisition.

#### THE ROLE OF THE PHONETIC REPRESENTATION

At this point we should explain in general terms how we view the role of a phonetic representation. It is characteristic of the perception of language, as A. M. Liberman, Mattingly, and Turvey (1972) have noted, that the perceiver remembers the gist of what was said and not the exact sentences, word for word, that the speaker uttered. That is, the speaker's original message is recalled in the form of a paraphrase. However, the paraphrase depends on the operation of a highly temporary memory system that contains a literal record of a small portion of the message as the hearer receives it. When we perceive a stretch of running speech, we rely on a working memory span of a few words that are held in phonological form. This can be demonstrated informally by abruptly interrupting a spoken communication. Typically, the listener can, on demand, repeat word for word the last few words or the last sentence that was uttered. One of the aspects of this short-term memory representation, then, is that it retains the most recent portion of the utterance in exact phonological form.

In speech, the primary function of this literal, limited-capacity and highly temporary memory representation, is, in Liberman, Mattingly and Turvey's (1972) view, to permit comprehension of the message. In order to comprehend what was said, we need to hold information about shorter segments (in our example, words) in memory until the meaning of the longer segments (here, sentences) can be grasped. But does reading necessarily require the same kind of memory representation as speech? If reading is rightly conceived as an alternative means of perception of language, then we may expect it to share many processes in common with the perception of speech. Reading involves interpretation of symbols that stand as surrogates for

speech segments. Thus, the reader's task, as we conceive it, is literally to convert print to speech, whether overtly, or (more usually in the case of the experienced reader) into some covert form. Although we do not rule out the possibility that read words can be held temporarily in some visual form, it seems reasonable to suppose that in reading, no less than in perception of speech by ear, the perceiver makes use of a phonetic representation in order to comprehend the message.

In the case of an alphabetic language, there is an additional reason for supposing that the reader derives a phonetic representation from print. The fundamental characteristic of alphabetic writing systems is that the letter symbols are a cipher on the phonemes of the language. Thus a reader who uses the alphabet analytically (in the sense of our discussion in the beginning of the previous section, *Linguistic Awareness and the Alphabet*) necessarily derives a phonetic representation. It is certainly the case that any reader must recode the written script phonetically if he is to decode a new word that he has never seen before. But does he need to recode phonetically words and phrases that he has read many times? Does he, in these cases, continue to construct a phonetic representation, or does he, as some believe, by-pass the phonetic level and go directly from visual shape to meaning?

It seems plausible to us that phonetic recoding might occur even with frequently read materials and may persist in experienced, skilled readers precisely, as we have intimated above, because a phonetic representation plays a functional role in comprehension. Elsewhere (Liberman et al., 1976; Shankweiler et al., 1976), we have speculated that the perceiver needs a phonetic base in order to index the mental lexicon and to reconstruct those prosodic cues so essential to comprehension of spoken language that are not directly represented in print.

Apart from our speculations on the role of a phonetic representation in reading, there is much experimental evidence that phonetic recoding does typically occur in a variety of situations in which the perceiver is confronted with visually presented linguistic material that he has later to recall. Most of the relevant experiments take the following form: lists of letters or alphabetically written words are presented to be read and remembered. Confusions in short-term memory are based not on visual similarity, but on phonetic similarity to the presented material. Conrad (1972) has noted that even nonlinguistic stimuli may be recoded into phonetic form and stored in that form in short-term memory. It was found, in this connection, that in recall of pictures of common objects, the confusions of children aged 6 and over were clearly based on the phonetic forms of the names of the objects, rather than on their visual or semantic characteristics.

All of these experiments are relevant to the assumption that even the skilled reader might recode phonetically in order to gain an advantage in short-term memory and to utilize the primary language processes he already has available to him.

In saying this, we do not imply that the only way to obtain meaning from script is via the intermediary of a phonetic representation. Our intent, rather, is to question the assumption (cf. Bever and Bower, 1966) that a direct mapping strategy that bypasses the phonetic level would always be the preferred mode of the mature reader. Such an assumption is unwarranted in

our view because it overlooks the large bulk of evidence that suggests that the organization of short-term memory is inherently phonetic.

#### Phonetic Coding in Good and Poor Readers

As we have seen earlier, a significant characteristic of the poor reader is his difficulty in identifying the phoneme, the unit most directly represented by the alphabet. In view of the short-term memory requirements of the reading task and evidence for the involvement of phonetic coding in short-term memory, we might expect to find that those beginning readers who are progressing well and those who are doing poorly might be further distinguished by the degree to which they rely on phonetic coding.

To explore the hypothesis that good and poor readers differ in the degree to which they use phonetic coding in short-term memory, we have carried out three sets of experiments with second graders.

In two of the experiments we used a procedure similar to one devised by Conrad (1972) for adult subjects in which the subject's performance is compared on recall of phonetically confusable (rhyming) and nonconfusable (nonrhyming) letters. In the first of these experiments (Liberman et al., 1976), the stimuli were strings of five uppercase consonants, half rhyming (drawn from the set BCDGPTVZ) and half nonrhyming (drawn from the set HKLQRSWY), presented tachistoscopically in a 3-sec exposure. Recall was tested under two conditions--immediately after presentation and after a 15-sec delay. In the second experiment (Shankweiler et al., 1976), the same procedure was followed except that the letters were presented auditorily on tape. Since auditory presentation requires successive input, a parallel condition using visual serial presentation was added in this experiment.

No matter whether the presentation was visual or auditory, simultaneous or successive, the results were virtually identical. Though the superior readers were better at recall on the nonconfusable items than were the poor readers, their advantage was virtually eliminated when the stimulus items were phonetically confusable. Though the effect was particularly marked in the delay condition, phonetic similarity always produced a greater penalizing impact on the superior readers than on the poor ones. It made practically no difference whether the items to be recalled were presented to the eye or to the ear.

These first two experiments strongly suggested that the difference between good and poor readers in phonetic coding will turn on their ability to use a phonetic representation, whatever the source, and not merely on their ability to recode from script. However, two major criticisms might be leveled at both experiments. First, since the stimuli used were strings of unrelated consonants, it is questionable that the results could be generalized to more realistic reading situations. Second, since the procedures did not control for the effects of rehearsal, the differences between the two reading groups might be accounted for by different rehearsal strategies.

A third experiment, using an adaptation of the Hyde and Jenkins recognition memory paradigm (Hyde and Jenkins, 1969), addressed itself to both criticisms (Mark, 1977). The subjects were given a list of 28 words to read aloud, followed by a second, or recognition list, containing all the



original words and, in addition, 28 new words, or foils. Half the foils were phonetically confusable with (but visually dissimilar from) a given word on the original list. The remaining 14 foils had no rhyming counterpart on the list. The subjects were required simply to respond yes or no as to whether a given word on the recognition list had appeared on the original list. Once again, though the stimuli were words, not strings of letters, and though rehearsal could not have been involved, the good readers were much more strongly penalized by the confusable items than the poor readers.

We regard these as interesting results. It is a relatively easy matter to demonstrate that good readers do better than poor readers on a variety of language-dependent tasks. In these three experiments, however, we have been able to show that it is possible to penalize good readers by making it disadvantageous for them to use a phonetic coding strategy. Therefore, it now seems reasonable to conclude with some confidence that good readers are more likely than poor readers to use a phonetic coding strategy effectively.

#### THE ANALYSIS OF READING ERRORS

One aim of our early research efforts (Shankweiler et al., 1972) was to determine whether the errors made by beginning readers, when they attempt to read words and syllables, pattern consistently and if so, whether an analysis of their pattern of errors might provide insights into the problems of reading acquisition. Accordingly, we carried out a phonetic analysis of reading errors in a number of experiments with beginning and disabled readers. A consistent pattern emerged from this analysis: errors on the final consonant of a CVC syllable were roughly double those on the initial consonant, while errors on the medial vowel exceeded those on consonants in both initial and final position. It seemed apparent from these experiments that the error frequency varied systematically with the position of the target phoneme in the syllable.

We considered that this distribution of errors in the syllable could be interpreted as a reflection of the child's lack of understanding of the phonological segmentation of his spoken language. If a child had not yet developed an ability to analyze the phonetic structure of his speech, he might be expected to show just this pattern of error--success with the initial segment, which can be extracted without further analysis of the internal structure of the syllable, and comparatively poor performance beyond that point. Such a child, who knew some letter-to-sound correspondences and also that he must scan in a left-to-right direction, might simply be searching his lexicon for word, any word, beginning with a phoneme that matches the initial letter. By this reasoning, if he were presented with the word big, he might, in context, give a response like beautiful, or out of context, butterfly. Neither response could occur if he were searching his lexicon, as he should, for a word that has three phoneme segments corresponding to the letter segments in the printed word. If, however, he is unaware that words in his lexicon have a phonetic structure, or if he has difficulty in determining what that structure is, his errors would increase after the initial segment. As to the relatively high incidence of his errors on vowels, it could, in these early experiments, have been simply attributed to the imbedded position of the letter representing the vowel in the CVC syllable of the stimulus list.

Though the pattern of consonant and vowel errors obtained in this early work was suggestive, certain controls were needed before we could accept its reliability. The purpose of a new series of experiments (Fowler, et al., 1976) was to confirm this pattern and, by the addition of various controls, to test its generality. Second, third, and fourth graders were the subjects in the new experiments. They were asked to read items from two lists of words<sup>1</sup>, one in which the incidence and location of the consonant phonemes were controlled, and the second in which these conditions were taken into account for the vowels.<sup>2</sup>

#### Differences in Consonant and Vowel Error Patterns

In this new set of more fully controlled experiments, we found the same pattern of consonant errors as previously obtained. Though the absolute number of errors decreased as the grade level of the child increased, the consonants in final position continued to produce approximately twice the number of errors as those in initial position. We were able to conclude that the consonant error pattern did indeed represent a true position effect and could, with some confidence, be attributed to the difficulties of phonological segmentation.

In contrast to the findings on consonant misreadings, errors in vowels showed no effect of position. When the vowels were placed in initial, medial, and final position, the errors did not vary systematically in frequency according to their location. Moreover, vowels continued to elicit a greater number of errors regardless of their location in the syllable. Thus, the high vowel error rate in the earlier experiment could no longer be explained by the medial position of the vowel and could not be related primarily to the difficulties of phonological segmentation.

The possibility that consonant and vowel errors might have different causes was supported by the results of a further analysis that took account not of the location of the errors in the syllable, but of the phonetic nature of the substitutions. In that analysis, it was found that consonant errors were systematically related to the target phoneme in the word, differing from it most often in only one of the three distinctive features of consonants (voicing, place of articulation, and manner of articulation). The proportion of consonant errors sharing two features with the target phoneme was remarkably stable across the grades: 60 percent of second-grade errors, 61 percent of third-grade errors, and 62 percent of fourth-grade errors. The results suggested, therefore, that phonetically motivated substitutions contribute substantially to the consonant error pattern both at the very early stages of reading acquisition and beyond. Vowel errors, in contrast, were

---

<sup>1</sup>A third list, used to study the error pattern in relation to the orthographic complexity of the vowels, was presented at the same time. Those results will be described in detail in a later paper.

<sup>2</sup>Ideally, it would have been desirable to provide both the consonant and vowel controls within one list. Contingencies relating to reading and vocabulary levels made this impossible to achieve.

not systematically related to the phonetic features of the presented vowel (tenseness, tongue advancement, tongue height, and diphthongization): indeed, the feature distribution of the vowel errors was essentially random at every grade level. Thus, the concept of featural similarity, so successful in rationalizing the substitutions among the consonants, does not enable us to understand the vowel errors.

The contrasting results obtained for consonants and vowels are indeed striking. The opposition of these phonetic classes was revealed, as we have seen, by both approaches to error analysis: the first, in which we investigated misreadings in relation to their location in the syllable, and the second, in which we considered the phonetic characteristics of the substitutions. As to the consonants, their position in the syllable accounted for the frequency of their occurrence, while the phonetic features of the target phoneme largely determined the nature of the substitution. With the vowels, on the other hand, neither of these relationships obtained. Factors other than these must, therefore, be considered to account for the vowel error pattern. One factor that suggests itself immediately is the variability of vowel orthography. Whereas the rules relating spelling to phonetic segment are relatively straightforward for consonants, they are quite complex for English vowels. Work in progress appears to single out the complexity of the orthography as a contributing factor in the vowel error pattern, though it was not so for the consonants.

#### The Error Pattern and Nonvisual Factors in Reading

These differences in error pattern lend credibility to the position taken by us and other investigators (Liberman, Shankweiler, Orlando, Harris, and Bell-Berti, 1971; Vellutino, Steger, and Kandel, 1972; Vellutino, Pruzek, Steger and Meshoulam, 1973) that visual perceptual factors are not sufficient to account for the difficulties of the beginning reader. It is hard to see how deficits in scanning, eye movements, and/or the discrimination of the optical form of letters can explain the differences we have found in consonant and vowel error patterns. Taken as optical shapes, the set of letters representing consonants is not marked in any distinctive way from the set representing vowels; the differences in error pattern between consonants and vowels, therefore, cannot be related to a classification based on visual characteristics.

Consonants and vowels do, in contrast, form distinctive categories in the language, with different functional roles in communication that might well lead to correspondingly different error patterns. Considered from the standpoint of their contribution to the phonological message, consonants carry the heavier information load. (A demonstration of this fact can be easily made: one needs only to compare the information obtainable in a sentence from which all the vowels have been deleted with one in which the consonants have been similarly treated.) The vowels, on the other hand, are the nucleus of the syllable structure and as such are the carriers of prosodic features. They are more subject than consonants to phonetic variation across individuals and dialect groups, and more subject to phonetic drift over time. As we suggested in an earlier paper (Shankweiler et al., 1972), the relatively greater variability of vowels than consonants may even account in part for the different ways these segments are represented in the orthography, particularly the larger variations in vowel spellings.

Additional evidence that language-related, rather than visual factors may be critical in early reading acquisition comes from a series of studies we have begun with second grade good and poor readers. In this recent research, we are investigating coding in short-term memory, not the error pattern in reading, but the results of one of the experiments are nonetheless directly relevant here. The paradigm used was an adaptation of Kimura's test of memory for recurring figures (Kimura, 1963). In this test, a series of stimuli are presented consecutively and the subject simply has to report yes or no as to whether the stimulus has already been seen in the series. There are four recurring stimuli that are exposed once in each set of 10 cards, randomly interspersed with six nonrecurring stimuli. Eight sets (of 10 cards each) make up a total of 80 cards in the test. The first set of 10 cards constitutes the presentation trials; the following seven sets are the recognition trials. This same procedure was carried out with three different sets of 80 stimuli--nonsense designs, photographed faces, and nonsense syllables. The results speak for themselves. The poor readers were slightly better than the good readers in memory for nonsense designs, but not significantly so. There was also no difference between the two groups of readers in face recognition. The good readers were better than the poor readers only in the nonsense syllables test, and there the difference was highly significant. Thus, despite identical procedures, neither nonlinguistic visual task differentiated between the good and poor readers, while the language-based visual task did. We would reason that in the nonsense syllable task, though not in the others, the good reader had a clear advantage: he could recode the information phonetically and thus hold it more efficiently in short-term memory.

At all events, perhaps the most general implication of these findings and those we have obtained in the error analysis, is that they again underscore the importance of nonvisual cognitive processes in reading and, specifically, those relating to language, such as awareness of phonological segmentation, phonetic recoding, and knowledge of the orthography.

#### IMPLICATIONS FOR INSTRUCTION

It has become fashionable to say that very little is known about how to teach reading and that the teacher makes a greater difference than the method. We would agree that the teacher's flexibility and wisdom in adapting existing curricula to meet individual differences, as well as his/her ability to recognize the necessity for doing so, will always be important variables in the success of any instruction procedure. However, we would also maintain that the little we do know about reading is often not reflected in reading curricula. If it were, even the less creative teacher might be more successful and the proportion of children resistant to reading instruction might be decreased.

To take a very basic example consider what we know about our writing system--namely, that it is alphabetic and not ideographic. From this, it would seem to follow that instructional procedures should inform the child early on that the printed word is a model of the component phonemes and their particular succession in the spoken word. Conversely, it would follow that the instruction should not, as it often does, mislead the child into assuming that the printed word is an ideographic symbol, a notion that will have to be corrected later, and, apparently for some children, with great difficulty.



Procedures that initiate the child into the mystique of reading by drawing his attention to the visual configuration ("remember this shape; it has a tail") and its associated meaning ("the one with the tail means monkey") without alerting him to the relevance of the sound structure of the word may lead the child into a blind alley. His ability to memorize the shapes and associated meanings of a handful of words may lull him and his parents into the comforting belief that he can read, but may leave him stranded at that stage, a functional illiterate with no keys to unlock new words.

Teaching a child how to use an alphabetic system to fullest advantage is complicated by the difficulty young children have in explicitly understanding the phonemic structure of their speech. As we have said, phonemic analysis is hard because of the encodedness of spoken speech into units of syllabic size; syllabic segmentation is demonstrably much easier. However, it need not follow that the phonemic level of analysis should be by-passed at the beginning in favor of the syllable or the word. Instead, perhaps the child can be given a better preparation for phoneme segmentation before reading instruction begins. With that preparation, certain elements of both the so-called phonic and syllabic methods can be introduced later to good effect.

#### How to Prepare the Child for Phonemic Analysis

The groundwork for this difficult level of analysis begins at home before the child is old enough to go to school. A proper foundation laid at this point can continue to be built upon in the pre-reading stages of kindergarten and at each succeeding stage of reading acquisition.

Word-play in Early Childhood. Games in early childhood that draw the child's attention to the phonemic content of his spoken language and that give him extended practice in "playing" with words may provide a foundation for future segmentation ability. Examples of such word play would include the learning of nursery rhymes, and the introduction of rhyming games that use both real words and nonsense syllables. The value of rhyming activity is that it varies the phonemic content while making few semantic or syntactic demands on the child.

Pre-Reading Techniques. When the child reaches kindergarten, pre-reading techniques would stress the phonemic structure of the spoken word before the written letters are introduced. "Listening games" that require the child to identify the initial, medial, and final phonemes in spoken words are in common use and need not be described here. Our only complaint with them is that they are not emphasized sufficiently in pre-reading training, and that, in actual practice, they often stop with the initial consonant. Teacher-devised methods that might help the child to hear sounds in words are limited only by the creativity of the teacher. One teacher<sup>3</sup> reports that she began prereading instruction for her kindergarteners at the Horace Mann School in New York by first teaching them to listen for the five short vowel sounds in words. Among the games she describes is one which seems particularly useful. In the first stage of the game, the teacher says a given vowel sound ("ă") once, twice ("ă ă"), or three times ("ă ă ă") and asks the class in each case to raise as many fingers as sounds they have heard. After the

<sup>3</sup>Marian Howard: personal communication.

children can do this correctly with all the short vowel sounds, she adds a consonant to the vowel, thus producing VC syllables ("am," "it," "op," etc.). She intersperses these syllables with single phonemes of the previous lesson and again asks for finger raising. She then progresses to consonant-vowel (CV) syllables, thence to CVC, CCVC, etc., varying vowels and consonants at each stage as needed. She reports that after instituting this "auditory program" in the fall, she could begin teaching reading by Christmas, and 90 percent of her kindergarteners were decoding print by April (the date of her report to us).

Several auditory training programs that emphasize the analysis of syllables into phonemes (rather than the discrimination of nonspeech sounds) have been available commercially for some time (see, for example, Lindamood and Lindamood, Auditory Discrimination in Depth, 1969), but none, to our knowledge, has as many worthwhile features as that outlined by the Soviet psychologist, Elkonin (1973).

In the procedure described by Elkonin, the child is presented with a line drawing of an object, animal, etc., the name of which is in his active vocabulary. Below the picture is a rectangle divided into sections equivalent to the number of phonemes in the pictured word. The child is taught to say the word slowly, putting a counter into the appropriate section of the diagram as he pronounces the word. After this "game" has been played with many different pictured words and the child can do the task successfully without the diagram, the idea of vowel and consonant sounds is introduced. At this time, the color of the counter is differentiated for the two phonetic classes--say, pink for the vowels, white for consonants. The child is first caught the difference between them with one vowel sound, being asked to put down a pink counter whenever he hears that sound. Not until the child can do this with the five short vowel sounds is the graphic form corresponding to the sound introduced.

The Soviet procedure has many pedagogical virtues. First, the line drawing keeps the whole word in front of the child throughout the process of analysis so that he does not have to rely on auditory memory to retain the word being studied. Second, the diagram provides the child with a linear visual-spatial structure to which he can relate the auditory-temporal sequence of the spoken word, thus reinforcing the key idea of the successive segmentation of the phonemic components of the word. Third, the sections of the diagram call the child's attention to the actual number of segments in the word, so that he does not resort to uninformed guessing. Fourth, the combination of drawing, diagram, and counters provide concrete materials that help to objectify the abstract ideas being represented. Fifth, the procedure affords the child an active part to play throughout. Finally, the color coding of the counters leads the child to appreciate the difference between vowels and consonants early in his schooling.

The actual content of the Elkonin procedure can, of course, be varied to fit the needs of the particular child or group of children, thus permitting its use not only for kindergarteners but also as a remedial technique for older children as well. The teacher can, for example, select for analysis syllables that contain whatever phonemes in whatever sequence she deems appropriate.

Three general rules might be suggested for the selection of syllables to be segmented. First, for this early training period, the noise portion of a fricative-like /s/ or the nasal murmur of /n/ or /m/ would be the consonants of choice for the prevocalic position in the syllables to be analyzed. These have the advantage that, unlike other consonants (particularly the stops), they can be produced in isolation. They can thus be used to acquaint the child with the general idea of word analysis without undue interference from coarticulation. Second, since two-segment analysis is easier than three-segment (Helfgott, 1976), training in segmentation might start with two-phoneme syllables. Finally, pilot data (Treiman, 1976) suggest that VC syllables are easier to analyze than CV and that both are (as we have said) easier than CVC syllables. Therefore, a vowel-consonant (VC) to CV to CVC succession in segmentation training would probably be most efficacious.

Another approach to training in phonological analysis, the elision technique outlined by Rosner (1975) in his "auditory skills program," places a somewhat greater conceptual burden on the child, but could profitably be used in conjunction with the Soviet procedure. It is always useful to offer a variety of different methods for attaining the same goals--with the proviso that the emphasis in the auditory training should be on the analysis of the sounds of speech. Training in nonspeech sounds, which are processed quite differently, cannot be expected to have the same effect (Liberman, 1971).

Once the child has been taught, by whatever method, to segment spoken syllables into their phonemic components, the graphic representations of the phonemes can be introduced. The Elkonin technique of adding the letter form to the blank counters might be adopted for teaching the graphic representation of the short vowels and one or two consonants. Thereafter, it would probably be preferable to shift to a more direct procedure for teaching the letters and their phonemic equivalents. This is the stage at which the child progresses from the prereading phase to actual reading instruction.

#### Basic Procedures for Initial Reading Instruction

We believe that the primary emphasis in teaching to read in an alphabetic system should be on mapping the components of the printed word to those of the spoken word. This analytic conversion from print to speech is best accomplished, in our view, by a method that presents reading, phonics, spelling, and handwriting in coordination with each other so that the instruction in each of these skills reinforces and illuminates the others. The integration of these four aspects of alphabetic communication serves to inform the child that they are indeed different facets of the same process and not separate, unrelated skills.

The First Step: Letter Names and Sounds. We would begin reading instruction, as many so-called phonics programs do, by teaching the child to associate the shape of the letter with its name and the sound it makes. We have come to agree with Mathews (1966) in his appraisal of this crucial first step: "... no matter how a child is taught to read, he comes sooner or later to the strait gate and the narrow way: he has to learn letters and the sounds for which they stand. There is no evidence whatever that he will ultimately do this better from at first not doing it all."

The simplest and most efficient way of teaching the sound-symbol correspondences is by the direct teaching of paired associates. The child should not be expected to abstract the correspondences for himself by a discovery method. Though some can do so, too many will fail. Useful materials for teaching the alphabet are alphabet cards that include not only the upper and lower case form of the letter, but also the mnemonic of a pictured key word beginning with the sound of the letter (Slingerland, 1971). On presentation of the card, the child is trained to recite the name of the letter, its keyword, and its sound (a, apple, ä). As the child learns each vowel, its symbol should be listed in a vertical column on the blackboard and reviewed each day. After the child has learned the five short vowels in this way, a few consonant symbols are introduced. Teaching of the remaining consonants by the same procedure can be continued in tandem with the next step. Meanwhile, the child is taught to write these same letters that he has learned to identify, not an unrelated series of letters presented in a separate "writing lesson."

Conversion from Speech to Print. The next step in most reading programs that emphasize phonics would probably be "blending." Since letter-to-sound correspondences have been learned in isolation, the traditional phonics method requires that these be combined or blended to form words. There the method runs afoul of the fact about speech that we have emphasized earlier: the spoken word is not a merging of a string of consecutive sounds. In speech, information about the three segments of the word "cat" is encoded into a single sound, the syllable. Therefore, no matter how fast the consecutive phonemes are spoken, "kuh-a-tuh" merged together consecutively will produce only the nonsense trisyllable "kuhatuh" and not the monosyllabic word "cat" (see Liberman, 1971 and A. M. Liberman et al., 1967 for extended discussions of this point).

How can we get around the problem of the fusion of phonemes by coarticulation? Though she also uses the more traditional blending method, Slingerland (1971) describes another technique which solves this fairly well. In effect, it is a spelling procedure that goes from speech to print and builds on skills that have been learned in the prereading program. Instead of demanding of the child the impossible task of blending "huh-a-mm" to produce "ham", the teacher first says the word, "ham", slowly, emphasizing the medial vowel. The child repeats the word, listens for the vowel sound, selects its letter card (color-coded as a vowel) from a wall pocket-chart and places it in a lower tier of the pocket-chart. The teacher then repeats the whole word and asks the child for the initial sound in the word. He selects the appropriate letter card, identifies it, and places it at the teacher's direction, in front of the vowel ("Where does it go? Before the a, because it's the first sound we hear"). The teacher then draws his finger along the two letters that the child has placed in the lower tier and says: "Now we have made 'ha'. Let's listen to our word again. Our word is 'ham' (drawing out the sounds). What is the last sound we hear in 'ham'? That's right, it's 'mm'. Find the letter that makes the 'mm' sound. Where do we put the m? At the end of the word, because it's the last sound we hear." The lesson continues with the child reading aloud the whole word that he has just constructed and ends with the child writing the word either on the blackboard or at his desk and reading it back after he has written it.



This procedure makes concrete for the child a key fact about writing that is difficult to explain in the abstract, namely, that temporal succession of the overlapping and nondiscrete speech segments (the phonemes) is represented spatially by a left-to-right linear succession of discrete characters (the letters).

A question that arises about this particular lesson is whether it might not confuse a child who has sequencing problems, since it requires him to start word analysis with the medial vowel sound and then to shift forward to the initial consonant sound. The answer is that, in actual practice, it does not seem to cause confusion. Typically, most children have sequencing problems in early reading acquisition only because they do not understand about the sound structure of the word and its relation to the written word. This spelling procedure has been preceded by much practice in listening for the components of spoken syllables. By building upon a foundation of knowledge of the sound structure of the word, the spelling procedure simply clarifies the relationship of the spoken word to print.

Thus far, the child has learned the letters and their sounds in isolation and has been taught, without using questionable blending methods, how to convert speech to sequences of letters, that is, how to analyze the spoken word and to construct its written model. But he still needs to be taught how to go from print to speech.

Conversion from Print to Speech. The next step is probably the most critical one since it should prepare the child to make the conversion from any printed word to speech, which is what early reading acquisition is all about. We are indebted to two teachers, Nancy Chapel and Cynthia Conway, learning disability specialists in the Greenwich, Connecticut public schools, for a sequence of lessons that has been highly successful at this stage of reading training.<sup>4</sup> Their procedures can be best characterized as a modification of the "linguistic" method of minimal contrasts, in which the unit under study is the syllable. The goal of their procedures is to make the conversion from printed syllables to speech more nearly automatic by circumventing the letter-by-letter sounding out and blending of the phonics method. The difference between their procedures and other syllabic methods is in the added structure built into the procedure that elucidates the internal construction of the syllable for the child.

In the Chapel-Conway lessons, the short vowels are listed on the blackboard in a vertical column and reviewed, just as they had been during the alphabet drills. At this time, however, a dash is added after each letter (a\_, e\_, i\_, o\_, u\_). The child is taught that the short vowel is always followed by a consonant and that the dash represents a missing consonant that will be filled in later. He is then taught the game of adding a letter in front of the short vowel and pronouncing the resultant combination (ma\_, me\_, mi\_, mo\_, mu\_). The prevocalic consonant is then varied (sa\_, se\_, si\_, so\_, su\_, etc.). Meanwhile, the children are encouraged to think of words beginning with those syllables and are taught to fill in the

---

<sup>4</sup>N. Chapel and C. Conway: personal communication.

missing final consonants in those words (man, met, mop, etc.). The lessons continue with the addition of consonant blends to the front of the vowel (smā\_, smē\_, smī\_, etc.).

When the short-vowel, closed syllable has been mastered, the idea of the long vowel is introduced, again with a structured model (a-e). It is pointed out that the missing letter in the model is now followed by an e, which is silent but marks the long vowel. Games of word construction with this model are then added. In the last stage, the child learns that when these consonant-vowel (CV) combinations appear alone without the added consonant (the dash representing the missing letter is now erased), the vowel is long and matches the letter name.

The child now has at his command a number of the major elements needed for decoding phonetically regular words. He can read closed syllables much more readily than he would if he had to depend on three-step (C-V-C) analysis and blending. At worst, since he knows CV syllables, he will have to resort only to a two-step blending (CV-C) that has been found to be easier (Helfgott, 1976). The basic contrast between the short and long vowels has been clarified, as well as that between closed and open syllables. Both of these understandings will be of importance to the child in learning to read polysyllabic words and words with more complex vowel orthography.

In conclusion, we must emphasize that we do not pretend to have developed a reading curriculum. What we have offered here are simply the outlines of a few basic procedures for initial reading instruction that seem to follow logically from what is known about the reading process, and that have proved successful in informal tests by teachers in the field. We would expect that the use of these and other procedures that relate print to speech will work more rapidly to achieve "reading for meaning," with fewer casualties, than could be accomplished by a program that stresses meaning at the outset.

#### A POSTSCRIPT ON THE CONTRIBUTION OF ORTHOGRAPHY TO READING PROBLEMS

As we have noted earlier, one source of difficulty in reading English is the nature of the orthography and the complex ways in which it represents the language. It is clear, however, that the complexities of the English orthography cannot be the sole explanation of reading difficulties, since some children continue to have problems even when the spelling of the words used in their instruction is phonetically regular and maps the sound directly (Savin, 1972). Nonetheless, we think it useful to look at early reading acquisition in an alphabetic writing system where the complications of orthography are minimized. Serbo-Croatian, the chief language of Yugoslavia, is such a case. The Serbo-Croatian writing system was devised on the principle of one letter shape for each phonemic unit in the language ("Write as you speak and read as it is written!" was the working motto of F. S. Karadić who introduced the new orthography).

However, before we can consider the consequences of the regularity of Serbo-Croatian orthography, we must take note of another characteristic of that writing system, namely that, for reasons of politics and religion, two

alphabets--one Cyrillic and the other Roman--were developed. Though they both represent the language quite directly, these two alphabets bear a complex relation to each other.<sup>5</sup> While some letters in the two alphabets share both the same shape and the same phonetic value, others are the same in shape but have different phonetic values. In still other instances, different letter shapes are used to represent the same phonetic units. Despite all these possibilities for confusion and interference, one of us was assured in a recent visit to Belgrade schools<sup>6</sup> that the double alphabet presents no problem: all the children learn the forms and letter-to-sound correspondences of both alphabets by the end of the second grade. The children are taught one alphabet for the first year and a half, and then master the other by the end of the second year. This should certainly give pause to those who would espouse visual-perceptual and simple memory deficits as causal factors in early reading disability--that is, if their faith had not already been somewhat shaken by the ability of Japanese first graders (and recently even kindergarteners) to learn the shapes and sound correspondences of two different sets of some 49 kana symbols (Makita, 1968).

As to the consequences for reading acquisition of the simple orthography of the Serbo-Croatian writing system, that is harder to evaluate. In the first place, children in Yugoslavia enter school at age seven, thus affording them an extra year of development before they must face the reading task. Second, no data are available on the actual incidence of reading disability.

It would appear, however, that some children do have reading problems, because the schools have developed extensive programs of prevention and remediation. One school we visited in Belgrade, for example, had a thorough

---

<sup>5</sup>Since Serbo-Croatian has two distinct alphabets for the same language, but with various overlaps in letter shapes and their correspondence, questions arise about how these ambiguous letter shapes are interpreted and where in the processing sequence the assignment to one alphabet or the other is made. Michael Turvey of Haskins Laboratories in collaboration with George Lukatela of the Department of Electrical Engineering at Belgrade University have begun a series of crosslanguage studies to investigate these interesting questions (Turvey: personal communication).

<sup>6</sup>We are grateful to Djordje Kostić, director of the Institute of Experimental Phonetics and Speech Pathology, for providing us with illuminating insights into the Serbo-Croatian language. Special thanks are due to Spasenija Vladislavljević of the Institute for arranging the school visits and serving as interpreter, guide, and informant throughout our discussions with teachers and school administrators. We are also particularly indebted to Ljubica Taipi, director of Branko Radicević school in Novi Beograd, for her generous cooperation in permitting us to talk freely with her staff, and to Ljubica Budinirović, vice-director, for her informative review of their educational programs. Numerous staff members there and in other schools in Belgrade also deserve grateful acknowledgement, but space does not permit mentioning them all by name.

preschool screening procedure. In the spring before school entrance, all the children are individually examined for intelligence, handedness, speech and motor development, socio-cultural background, and emotional adjustment. Those with special problems are identified and given additional diagnostic testing and assistance as needed. Another facet of the built-in preventive program in the primary grades of this school is team teaching. Teachers of each grade exchange classes at frequent intervals throughout the school year and hold regular consultations with each other on how best to teach all their problem children; if they decide that additional special remediation in reading is indicated, they refer the children to therapists who advise the teachers and work directly with the children.

It is interesting to note that the basic training of these therapists is in phonetics and speech pathology. We should suppose that the educators require that background in their therapists because they assume a close relation between speech and reading. In any event, the therapy certainly reflects that particular bias, just as ours does. For example, heavy emphasis is placed in both developmental and remedial instruction on pre-reading drill and exercises in the analysis of the spoken word. Moreover, once the alphabetic letters are introduced, the procedures are again quite similar in general approach to those we have outlined here. That is, the instruction is directed toward clarifying for the child the relationship between the spoken word and its written counterpart.

The importance that Yugoslavian instructors attach to relating print and speech was made clear to us by Professor Spasenja Vladislavljevic of the Institute for Experimental Phonetics and Speech Pathology at Belgrade University, who is in charge of the training and supervision of the therapists. She illustrated her point by describing a typical first-grade reading lesson that follows much practice in listening for sounds in words. The teacher pronounces the sound of the initial consonant CVC word to be read (always a nasal or fricative in the early lessons) and writes its letter on the blackboard with a line following it. As she draws out the spoken word for varying periods of time, she shortens or lengthens the line following the letter (s-----, s--, s-----). This exercise is repeated with the vowel (a-----, a-----, a--). Then both sounds are spoken and the interval between them varied and represented accordingly (s-----, a-----, s--a-----, sa---). The consonant in final position is then added (sa----t) and the word is spoken as a whole (sat). Finally, the word is written without the lines and read aloud. In subsequent lessons, the child is taught to read and write other words by the same procedures. When he has mastered a word, he writes it in his notebook and perhaps uses it in a written sentence that he also reads aloud. Thus, reading, writing, and spelling exercises are always coordinated, as we have also proposed.

In summary, it must be said that despite the regularity of the Serbo-Croatian orthography, some children--we do not know how many--apparently do encounter difficulties in early reading acquisition. What proportion of these ultimately become fully literate, we also do not know. We have no hard data on either of these questions, though we are told that in the end the children do well and reading disability is not a problem. At all events, crossnational assessments of reading achievement are difficult to evaluate.



In this particular case, one does not know how much weight should be given to the regularity of the orthography and how much to the special characteristics of the reading instruction. The answer to this question must await further research.

#### REFERENCES

- Bever, T. G. and T. G. Bower. (1966) How to read without listening. Project Literacy Reports No. 6, 13-25.
- Calfee, R., R. Chapman, and R. Venezky. (1972) How a child needs to think to learn to read. In Cognition in Learning and Memory, ed. by L. W. Gregg. (New York: Wiley).
- Conrad, R. (1972) Speech and Reading. In Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).
- Elkonin, D. B. (1973) U.S.S.R. In Comparative Reading, ed. by J. Downing. (New York: Macmillan).
- Fowler, C. A., I. Y. Liberman, and D. Shankweiler. (1976) On interpreting the error pattern in beginning reading. Haskins Laboratories Status Report on Speech Research SR-45/46, 17-28.
- Gallistel, E. and K. Ellis. (1974) Gallistel-Ellis Test of Coding Skills. (Hamden, Connecticut: Montage Press).
- Gibson, E. J. and H. Levin. (1975) The Psychology of Reading. (Cambridge, Mass.: MIT Press).
- Gleitman, L. R. and P. Rozin. (1973) Teaching reading by use of a syllabary. Read. Res. Quart. 8, 447-483.
- Helfgott, J. (1976) Phonemic segmentation and blending skills of kindergarten children: Implications for beginning reading acquisition. Contemporary Educational Psychology 1 (2), 157-169.
- Hyde, T. S. and J. J. Jenkins. (1969) Differential effects of incidental tasks on the organization of recall of a test of highly associated words. J. Exper. Psychol. 82, 472-481.
- Jastak, J., S. W. Bijou, and S. R. Jastak. (1965) Wide Range Achievement Test. (Wilmington, Delaware: Guidance Associates).
- Kimura, D. (1963) Right temporal-lobe damage. Arch. Neurol. 8, 264-271.
- Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
- Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington: Winston).
- Liberman, I. Y. (1971) Basic research in speech and lateralization of language: Some implications for reading disability. Bull Orton Soc. 21, 71-87.
- Liberman, I. Y. (1973) Segmentation of the spoken word and reading acquisition. Bull Orton Soc. 23, 65-77.
- Liberman, I. Y., D. Shankweiler, F. W. Fischer, and B. Carter. (1974) Explicit syllable and phoneme segmentation in the young child. J. Exp. Child Psychol. 18, 201-212.
- Liberman, I. Y., D. Shankweiler, A. M. Liberman, C. Fowler, and F. W. Fischer. (1976) Phonetic segmentation and recoding in the beginning reader. In Reading: Theory and Practice, ed. by A. S. Reber and D. Scarborough. (Hillsdale, New Jersey: Erlbaum Associates).

- Liberman, I. Y., D. Shankweiler, C. Orlando, K. S. Harris, and F. Bell-Berti. (1971) Letter confusions and reversals of sequence in the beginning reader: Implications for Orton's theory of developmental dyslexia. Cortex 7, 127-142.
- Lindamood, C. H. and P. C. Lindamood. (1969) Auditory Discrimination in Depth. (Boston: Teaching Resources).
- Makita, K. (1968) The rarity of reading disability in Japanese children. Am. J. Orthopsychiat. 38, 599-614.
- Mark, L. (1977) Phonetic recoding and reading difficulty in beginning readers. Unpublished master's thesis, Department of Psychology, University of Connecticut.
- Mathews, M. (1966) Teaching to Read Historically Considered. (Chicago: University of Chicago Press).
- Mattingly, Ignatius G. (1972) Reading, the linguistic process, and linguistic awareness. In Language by Ear and By Eye: The Relationships Between Speech and Reading. (Cambridge, Mass: MIT Press).
- Read, C. (1971) Pre-school children's knowledge of English phonology. Harvard Educ. Rev. 41, 1-34.
- Rosner, J. (1975) Helping Children Overcome Learning Disabilities. (New York: Walker and Company).
- Rosner, J. and D. P. Simon. (1971) The auditory analysis test: An initial report. J. Learn. Dis. 4, 40-48.
- Rozin, P. and L. R. Gleitman. (1976) The structure and acquisition of reading. In Reading: Theory and Practice, ed. by A. S. Reber and D. Scarborough. (Hillsdale, New Jersey: Erlbaum Associates).
- Savin, H. B. (1972) What the child knows about speech when he starts to learn to read. In Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).
- Shankweiler, D. and I. Y. Liberman. (1972) Misreading: A search for causes. In Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).
- Shankweiler, D. and I. Y. Liberman. (1976) Exploring the relations between reading and speech. In Neuropsychology of Learning Disorders: Theoretical Approaches, ed. by R. M. Knight and D. J. Bakker. (Baltimore: University Park Press).
- Slingerland, B. H. (1971) A Multi-Sensory Approach to Language Arts for Specific Language Disability Children: A Guide for Primary Teachers. (Cambridge, Mass.: Educators Publishing Service, Inc.)
- Treiman, R. A. (1976) Children's ability to segment speech into syllables and phonemes as related to their reading ability. Unpublished manuscript, Department of Psychology, Yale University.
- Vellutino, F. R., J. A. Steger, and G. Kandel. (1972) Reading disability: An investigation of the perceptual deficit hypothesis. Cortex 8, 106-118.
- Vellutino, F. R., R. M. Pruzek, J. A. Steger, and U. Meshoulam. (1973) Immediate visual recall in poor and normal readers as a function of orthographic-linguistic familiarity. Cortex 9, 368-384.
- Zifcak, M. (1976) Phonological awareness and reading acquisition in first grade children. Unpublished doctoral dissertation, University of Connecticut.

## Visual Processing and Short-Term Memory\*

M. T. Turvey†

### ABSTRACT

This paper is divided into two parts. The first erects the framework of indirect realism: the historical and theoretical backdrop against which the contemporary analysis of visual processing is conducted. In addition to setting the stage, the first part introduces and presents a thumbnail sketch of the principal characters, namely, those mechanisms which have assumed a central role in current interpretations of visual processing. The responsibility of elaborating on the personalities and capabilities of several of these principal characters rests with the second and larger part; to this purpose we describe in elementary--but it is hoped, sufficient detail--the methodology, findings, and intuitions that bear on these mechanisms of processing.

### PART A

#### Visual Information Processing: A preliminary portrayal

Ostensibly, the task before the visual information-processing theorist is to chart the flow of visual information within the human observer. The enterprise begins with the realization that visual experience is not an instantaneous reaction to optical pattern; on the contrary, there is an appreciable amount of time between the visual experience, and the occurrence at the eyes of the stimulation relating to a given aspect. Experimental observation is ready witness to this claim: a display exposed briefly to an observer and followed tens of milliseconds later by another display may be phenomenally obscured, or at least not identifiable. How might we then characterize the processes underlying perception--processes that appear to be temporally extensive? It is customary to adopt the point of view that the processes are hierarchical; they are spoken of as a succession of stages, of both storage and transformation, which map the structured energy at the receptors onto increasingly more abstract representations.

Taking the above characterization as our departure point, let us proceed to sketch the form of a visual information-processing system. The sketch conveys the gist of a number of separate but closely cognate portrayals (cf.

---

\*To appear in Handbook of Learning and Cognitive Processes Vol. V, ed. by W. K. Estes, (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).

†Also, University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

Neisser, 1967; Sperling, 1967; Broadbent, 1971); and though there are reasons to question the depicted system's worth (Turvey, 1975; in press), it will serve the present purposes well. The system sketched here has motivated a great deal of the research that will be of concern to the present chapter.

Pivotal to many accounts of the flow of information, is the idea of a brief, large-capacity literal memory interfacing the pattern at the receptor surface with the procedures responsible for that pattern's eventual identification. It is assumed that this memory preserves physical information about the pattern, in the sense that it is a precategorical or nonabstract representation. After Neisser (1967), this transient memorandum is referred to as iconic storage.

Two closely related operations performed upon the icon can be identified. One is selective attendance to some aspects of the store: inasmuch as the icon is a large capacity buffer, and inasmuch as subsequent mechanisms are thought to have a less generous capacity, then a process of selection is mandatory. The other operation determines the identity or class of the selected information. It is far better, perhaps, to say that what is determined is a description of the selected information that bids fair for subsequent, more durable memory, and even more importantly, provides a suitable basis for responding. In any event, the description given to the selected information is the responsibility of processes that are part and parcel of the observer's long-term model of the world. Short-term or primary memory is the immediate storage medium for the categorical description that results from the iconic storage/long-term model interaction. We need note now only a process by which the contents of the short-term memory are preserved on the one hand, and are woven into the fabric of permanent memory on the other--a process customarily dubbed "rehearsal"--and our elementary sketch of the visual information-processing system is complete (see Figure 1).

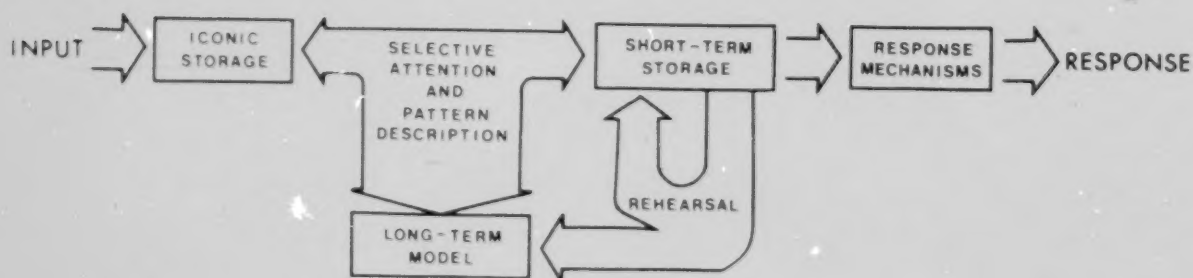


Figure 1: A typical visual information-processing scheme.



In the context of the above remarks comes the colorful phrase "flow of information," that is described as follows: at a given stage, one state out of a set of possible states is occasioned in very large part, by the occurrence of one of a number of possible states from an earlier stage (Broadbent, 1971). To elaborate upon the above model, first consider a coarse but instructive partitioning of the information flow by Broadbent (1971), and second, a more detailed apportionment from investigators in the field of artificial intelligence.

In Broadbent's (1971) view, there are essentially three loci in the flow of information where an ensemble of possible states can exist. At any particular moment in time, a particular state occurs as one of many possible states of the environment. It is this state that we commonly refer to as "stimulus."

On the supposition that the proximal stimulus (the optical arrangement on the retina due to the environmental state or distal stimulus), is imperfectly or ambiguously related to the environmental state, then more than one preliminary description of the stimulus is possible. The particular preliminary description that does occur is the "evidence" for the pattern-recognition devices, and based on this evidence, one of a number of possible outputs is selected--the "category state." It is supposed, of course, that there are rules relating the three classes of states: rules that map stimulus states to evidence states, evidence states to category states, and stimulus states to category states. To modulate the flow of information is to manipulate these rules. In illustration, imagine a man or woman looking at a display of intermixed letters and digits, with some of the letters and digits colored red, some colored green, and some colored yellow. If the observer is asked to report the red items, then he/she adjusts the rules linking stimulus states to evidence states; the ascribing of states of evidence is biased toward red items rather than yellow or green items. Suppose the observer is asked to report the digits in the display, regardless of color. That would be asking, in essence, that the rules relating evidence states to category states be adjusted: one evidence state for each of the stimulus states will be considered, but only those relating to the set of digits will lead to an output. Selection in the former case is said to be by "filter-setting," and in the latter case by "pigeon-hole setting" (Broadbent, 1971). In filter-setting, the source of the stimuli controlling the response is specified, but not the response vocabulary. In pigeon-hole setting, it is the vocabulary of responses that is specified, but not the source of stimuli. While the rules relating stimuli to evidence and those relating evidence to categories are reasonably manipulable, those that relate stimuli to categories are not so pliable. "Category-setting" is a gradual process contingent upon lengthy experience.

Let us now turn to the attempts in artificial intelligence research to contrive a machine that "sees" scenes in the sense that it describes an environmental outlay much as a human observer might. For example, what can you say about the scene--an arrangement of polyhedra--depicted in Figure 2? A truncated inventory of your responses would include comments such as: "there are five separate objects present;" "the cube-like object is supported by the rectangular-like object;" "the wedge is closer (to me) than the cube;"

and so on. What is evident is your ability to describe, with some aplomb, the three-dimensional arrangement represented by the picture, and to describe that arrangement in terms of a number of extremely involved relationships. How could we put together a device that would arrive at a similar description?

---

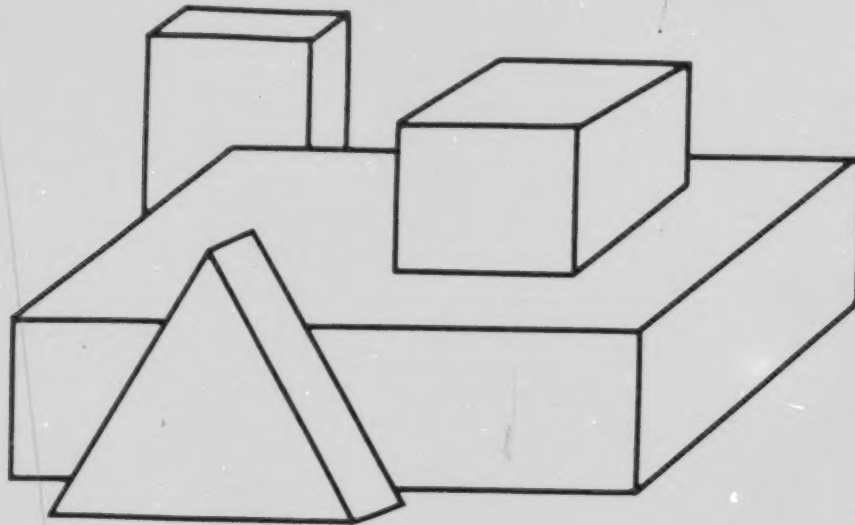


Figure 2: A scene of polyhedra.

---

In order to be clear in this enterprise, the following boundary conditions should be identified: first, all of the scenes the machine will be required to "see" are of variously arranged, opaque polyhedra; second, the eye of the machine is a TV camera and thus the pictures transmitted are defined by the grayness value at 1024 x 1024 different locations. It is now necessary to identify our machine's tasks and in so doing to recognize the parallel task of the human observer who, it can be argued, must begin with (like the machine) an array of points--the retinal mosaic.

Essentially, the recovery of the three-dimensional description from the array of points involves the serial construction of progressively more abstract representations. For the purposes of computation, a representation is defined as the specification of relations among a set of entities having certain attributes. The kinds of representations intervening between input and final description are suggested in Figure 3.

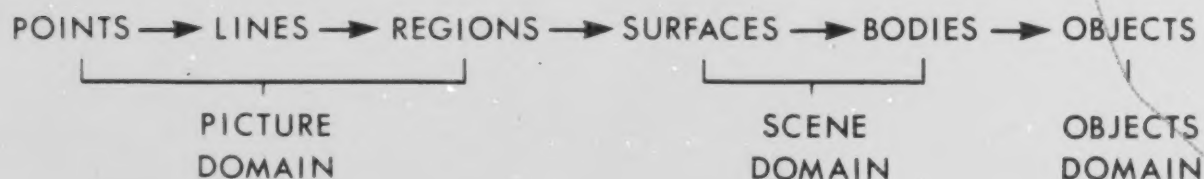


Figure 3: Scene analysis as a sequence of progressively more abstract representations.

---

Consider two examples of representations and what they entail: the regions representation and the surfaces representation. In the regions representation, the entities are regions, vertices, and boundaries; these entities have the property of shape and they relate in terms of "inside of" and "adjacent to." For the surfaces representation, the entities are surfaces, edges, corners, and shadows; their attributes are shape, slant, and albedo, and their relations are convex, concave, behind, and connected (Sutherland, 1973).

Suppose that the machine has already computed a description of a picture at the level of the regions representation. (A region is an array of points closed off by a set of lines cyclically arranged with coinciding end points.) The problem is to decide which of the regions go together, that is to say, how to segregate the picture into separate bodies. One way to solve this connectedness problem is to examine the implications of the vertices bounding the regions. Each vertex type provides evidence about the groupings of regions, and therefore can be exploited to specify links among regions.

Some of the more important vertex types are given in Figure 4. The arrow type of vertex is commonly caused by an exterior corner of an object where two of its plane surfaces form an edge. Thus, an arrow vertex implicates a link between the two regions that meet on the shaft of the arrow, but not between those that meet at its barbs (see Figure 5). On the other hand, a fork vertex permits the linking of all its bounded regions, for a vertex of this kind is due usually to the corner formed by three planes of one body (see Figure 5). In short, knowledge about the structure of bodies in the scene domain (Figure 2) allows humans, the constructors of seeing



ARROW



TEE



FORK

Figure 4: A few examples of vertices and the links they imply.

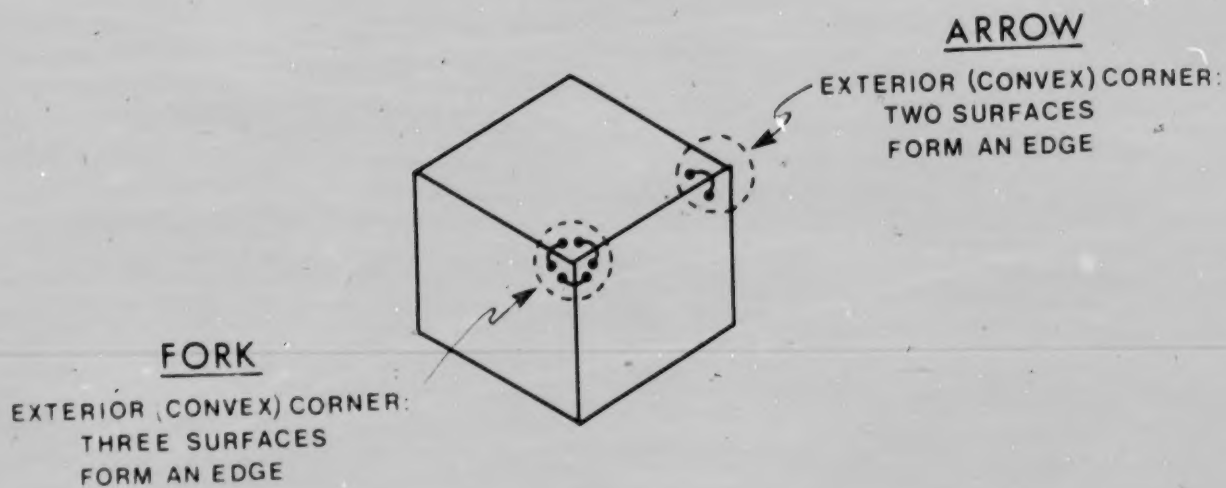


Figure 5: Vertices and corners.



machines, to adduce three-dimensional properties of vertices that are, of course, two-dimensional entities in the picture domain. A familiar program that uses these implications of vertices to map from the regions' representation to the bodies' representation, is that of Guzman (1969). Without any further detail, it is noted, by way of summary, that the Guzman program can successfully segment the scene, that is, specify the number of bodies, as depicted in Figure 2. It is also noted, however, that the program is error prone; it will treat holes as bodies and it will compute various other kind of incorrect segmentations. These failings impel the question: What other knowledge is needed in order that the program may operate more efficiently and provide, reliably, a more accurate description of scene? One solution, suggested by Clowes (1971), is to introduce knowledge about the permissible articulations of polyhedra surface in three-space. Within the Guzman interpretation schema for vertices, it was asked only how three-dimensional bodies project into the picture domain, and that was the limit on the higher-order knowledge embodied by the program.

Consider once again the fork vertex. All three regions can be linked if, and only if, the vertex actually specifies a convex corner in the scene domain. As Figure 6 shows, a fork vertex could be due to a concave corner, in which case it would be incorrect to link all adjacent regions. So, the mapping from vertex to corner is equivocal. This ambiguity can be circumvented in the Guzman program by using a procedure that looks at how many links exist between one region and another. When two regions are connected by only a single link, the link is deleted; in contrast, Clowes' (1971) solution to the ambiguity problem is to give his program a particular kind of knowledge about the world--precisely, given any two corners, the edge joining those corners must be invariant (for example, concave or convex) throughout its length. The systematic use of this fact about the relations among surfaces in three-space permits the program to arrive at sets of compatible corner interpretations.

At all events, the interpretation of local aspects of a picture is facilitated by nonlocal information. Inasmuch as Clowes' program may be adjudged superior to that of Guzman's (Sutherland, 1973), it appears that increasing the program's knowledge about the world enhances the program's capability to perform scene analysis.

A general conclusion from the research into computer perception is that highly flexible programs that use information and procedures in higher domains to reinterpret information in lower domains, are necessary to the construction of successful seeing machines. This conclusion merits serious consideration on two grounds: first, it suggests that hierarchical schemes that unidirectionally map from lower to higher representations are insufficiently powerful; and second, it implies that the major constraint on perceptual achievement is the highest domain in which knowledge about the world is available. With respect to the latter, it is supposed that a seeing machine that has a priori knowledge about the world at the level of the surface representation--but no higher--is less capable than one that instantiates knowledge about objects. If we assume that the objects domain is the highest domain, then we must conclude that the degree to which a perceptual device incorporates knowledge about objects determines the degree

of success that device will have in perceiving the world. As Gregory (1970) comments: "Perception must, it seems, be a matter of seeing the present with stored objects from the past" (p. 36).

---

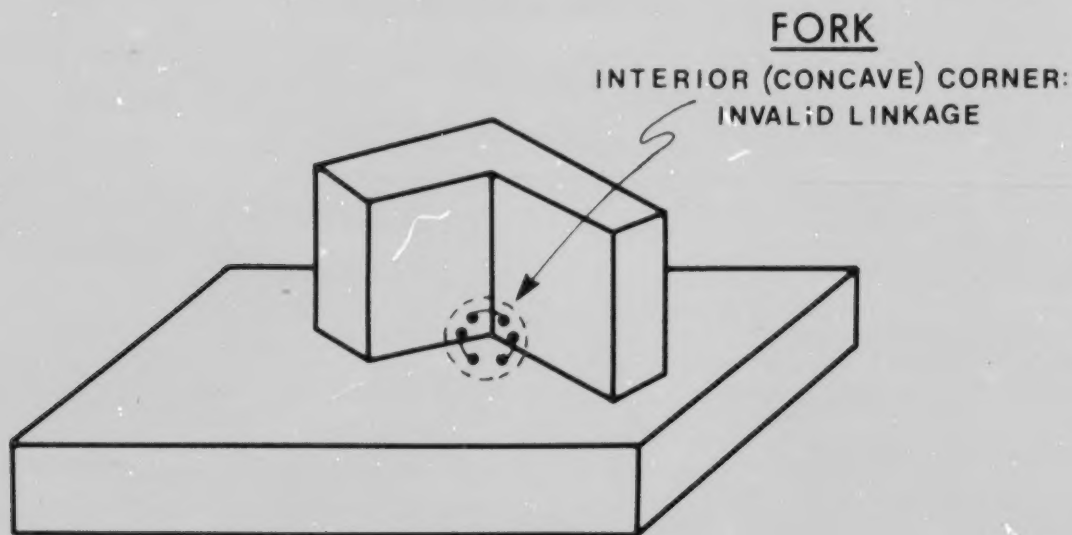


Figure 6: A fork vertex is equivocal.

---

#### Visual Information-Processing in Historical Perspective

Although a number of variously described schemes of visual information processing can be identified, and although it is the case that some are more formal and detailed than others, it is fair to comment that they are all tokens of the same type. The purpose in this preliminary section is to identify the assumptions most common to models of visual information-processing and to appreciate their historical consistency and ubiquity.

We owe to the nineteenth century scholar Helmholtz the exposition of the problem of perception as it is most commonly understood. As will be seen, there is little in the contemporary "flow of information" conception that was not touched upon by Helmholtz (1925) or his intellectual predecessors, among whom we may recognize Berkeley and Alhazan, whose insights were born before the end of the first millenium A.D.

It is assumed that the reader is familiar with the understanding that environmental states--distal stimuli--cannot determine visual perception,

since those states do not affect the sense organs directly; only the proximal stimulus--say, the retinal image engendered by a given environmental state--can do that. Unfortunately, orthodoxy holds that the proximal stimulus relates equivocally to the environmental state. For example, more than one image relates to a given object and two different objects may project the same image. The traditional argument runs that the proximal stimulus cannot likewise determine perceptipn; it can serve only as a suggestion, a clue if you wish, about the world's condition. Even on this point one must be more cautious, for in the view of many scholars, the proximal stimulus in vision is not so much a fleshed-out form, as the phrase "retinal image" suggests, as it is a mosaic of elements, or points of light that can have but two characteristics: intensity and wavelength.

Helmholtz's tutor, Johannes Muller, had concluded that the various qualities of conscious experience that we associate with the separate senses were due to specialized receptor-nerve systems. Each of these systems was selectively sensitive to a particular kind of proximal stimulation--the auditory receptor-nerve system to pressure waves, the visual receptor-nerve system to light--although each would, when artificially stimulated, yield an invariant response and an invariant quality of experience.

For Helmholtz, the worth of Muller's contention lay in the fact that one could account for the different qualities of conscious experience within a modality along very much the same lines: for each discriminable quality within a sense, there was a corresponding receptor-nerve system. Such being the case in a sense-system, any proximal stimulus relating to a given environmental state could be described as a collection of the responses of these specialized systems. In short, the proximal stimulus was said to be mapped into a set of primitive experiences, the so-called sensations.

It was remarked earlier that a point of light has only the dimensions of intensity and wavelength. Assuming that we can identify corresponding specific nerve energies, then the first stage of perception is a representation of the proximal stimulus as a patchwork of variously bright colors. Inasmuch as specific nerve energies for environmental properties such as solidity, size, distance, substance, were not discovered--indeed, not expected--then the task of the observer as conceived by Helmholtz was manifestly plain: to construct or infer these environmental properties from the patchwork of colors.

The assertion that perceptual experience is fashioned from a finite set of elemental experiences is common to contemporary visual information-processing models. However, the modern version takes a liberalized view of specific nerve energies. Benefitting from the neurophysiology of the times (for example, Maturana, Lettvin, McCulloch, and Pitts, 1960; Hubel and Weisel, 1967, 1968), it proposes that specific nerve energies do not simply exist for punctate stimulation; they can be said to exist for spatial optical arrangements such as those that correspond to contour character, convexity and differently oriented lines. According to Helmholtz's theory, these relational features would have to be carved out of the patchwork of variously bright colors through associative memory. In current theory they are given directly in the same sense that the points of color in the Helmholtz version

are given directly, that is to say, they are not mediated by memorial or intellectual problem-solving processes.

In any event, armed with our modern version of specific nerve energies, it might be tempting to argue that perception is just a matter of listing or combining features. However, perception founded upon features should be no more simple--no less devious--than perception based upon color patches. As has already been commented, the same proximal stimulus can stand for several environmental states. If proximal stimulation is inherently equivocal, then the first internal representation of that stimulation, whether it be in terms of color sensations or relational features, must be likewise.

Helmholtz, and long before him, Alhazan, supposed that we arrive at our impression of the state of the environment through a process of calculation--of logical inference--performed upon the elemental givens. In brief, a human's experience of an environmental state is an unconscious conclusion that is dependent upon his/her knowledge about the world. For Helmholtz, perception was most obviously a process over time involving procedures akin to testing and experimenting; further, since knowledge about the world was gleaned through experience, perception for him could not be divorced from memories. More precisely, in this view, perception is wrought by conception. The kernel of this position is given in Hochberg's (1974) eloquent paraphrase of Helmholtz: "...we perceive the object, scene or event which would normally fit the proximal stimulus distribution " (p. 21, author's italics).

It is evident from the above that the principal assumptions of contemporary visual information processing relate closely to those assumptions that formed the foundation of Helmholtz's system, namely, the existence of a finite set of primitive elements, the existence of procedures that could make infinite use of this finite set, the dependence of current perception upon stored knowledge about the world, and the temporal extensiveness of perception, that is, perception as a process over time.

A further aspect of Helmholtz's scheme deserves our attention. In order to reach a conclusion about a local point of light, neighboring points of light may have to be considered in the course of perceiving. Similarly, in order to reach a conclusion about an object property, such as size, other properties of the environment--for example, distance--have to be taken into account. One is reminded of nonlocal processing that is needed to disambiguate local interpretations in the determination of accurate scene descriptions by computer. Consider that the ocular equipment of humans and animals--and indeed--of machines, is mobile. The movements of the eyes, the head and the body result in larger and larger samplings of the environment over longer and longer time periods. If nonlocal aspects are necessary to the veridical interpretation of local aspects, one should ask: How nonlocal is nonlocal?

The puzzle is how to define the size spatially and temporally of the context in which the proposed inferential processes fashioning perception operate. One could argue that visual perception--visual information processing--operates on temporally discrete samples or moments of some fixed or variable size, that is to say, that it operates discontinuously in time on



the proximal stimulation. On the other hand, one might suppose that perception operates on a temporally continuous running sample of fixed or variable size with the perception at any instant being some function of the proximal stimulation sampled over the previous  $x$  seconds. These two notions may be labeled, respectively, the discrete moment hypothesis and the traveling moment hypothesis (cf. Allport, 1968). Tradition has favored the former. Thus, the scan of the eyes, a succession of fixations, is likened to a succession of discontinuous "retinal snapshots" (Neisser, 1967). Each retinal snapshot as a frozen slice of time presents a proximal distribution for interpretation with recourse to conceptual knowledge; the succession of perceptual conclusions so produced must then be integrated to achieve a reliable perception of the panorama. Often, a short-term storage mechanism is proposed as the medium permitting the possibility of this integration. At all events, the issue of perceptual moment and context sample size cannot be ignored, and it will be encountered in various forms in the following remarks.

#### Visual Processing as a Version of Indirect Realism

In the preliminary sections, some common variants of the contemporary approach to visual processing have been perused. The genesis of the approach has been given a brief historical look. By way of summary, it is manifestly obvious that visual information-processing ascribes to that version of realism that bears the epithet "indirect." By the term realism, it is asserted that there is an objective world, external to ourselves, that can be perceptually experienced; and by the term indirect, it is asserted that our experience of that world is secondhand, that is to say, it is mediated by a representation of the world.

Indirect realism in its most common form denies that the world "is the proximal cause of our perceptual experience" (Shaw and Bransford, 1976). In the chain of causes-and-effects from distal object to percept, the weak link is said to be the world, because the world, it is thought, does not structure the light distribution at the eye in a way that is specific to its properties. Hence, other sources of knowledge must come to the aid of the visual-processing system in order to enrich the impoverished perceptual information, and thereby insure the veridicality of perception. For indirect realism, therefore, our perceptual experience is not of the world, but of one of the other, and presumably later, links in the causal chain of perception. We can thus regard visual processing research as an enterprise that seeks to disclose the causal chain of perception. Some of the intuited links of that causal chain are examined in the second part of this paper (Part B).

#### PART B

##### Rationalizing Iconic Storage

In the visual information-processing scheme outlined at the outset and represented by Figure 1, the scheme is simple: the perceiver's long-term

model of the world intervenes between two hypothesized descriptions of the proximal stimulus, a briefly existing literal description, and a far more durable categorical description. The former is iconic storage; the latter is short-term storage. The techniques for revealing the brief icon will soon be highlighted. As a preliminary, though, the icon's raison d'etre--the motivation for speculating on the existence of a precategorical visual memory--is sought.

Broadbent's (1958) vintage description of the flow of information claimed that the mechanism responsible for category determination was of limited capacity and, moreover, that this mechanism was amodal. Consequently, when information arrived that exceeded the capacity, there had to be a filtering process that protected the categorizing mechanism from overload. Since the limited capacity device might become available without too much delay, it would be advantageous for the system to have at its disposal a mechanism for maintaining the prohibited information. If such was the case, the limited capacity mechanism, within limits, could delay attending to a source of information with relative impunity. In short, the assumption of a limited capacity pattern recognizer encouraged the notion of a precategorical buffer.

There is another source of encouragement for the idea of a buffer memory that derives similarly from a consideration of how pattern recognition operates. In this case, the pattern recognition device is of the analysis-by-synthesis type. Though originally popular in the realm of speech perception models (Halle and Stevens, 1962), analysis-by-synthesis was to become, in the hands of Neisser (1967), a provocative account of many visual phenomena, both common and exotic. In this view, identification proceeds by matching two descriptions, that resulting from a preliminary but far from complete analysis of the input, and that generated or constructed from the long-term model of the world. In this active matching process, the pattern-recognizer achieves identification of a given input, by questioning what is done to the patterns stored in long-term memory to make one of them look like one yielded by the preliminary analysis. Obviously, to answer this question, more than one comparison, on the average, will be needed. The system will have to hunt for a match, and while it does so, the preliminary analysis of the input must be preserved. From this point of view, a precategorical buffer is a necessity.

Both of the above are reasonably contemporary inspirations for proposing a literal sensory memory; both, however, echo more traditional concerns. It has long been known that the transduction from light energy to neural activity occurs rapidly, far more rapidly, it was thought, than the processes responsible for perception. A traditional argument of considerable currency (if one substitutes "features" for "sensations") is that sensations are produced with dispatch, but perception occurs at a more leisurely pace. For any sensation-based theory of perception, the existence of a high-capacity sensory store intervening between sensory processes and perceptual processes would appear to be mandatory.

There are other imaginable reasons for proposing the existence of a brief sensory memory, but the above will suffice for our present purposes.

It is time to proceed to the methodological problem of how to capture this fleeting memorandum.

### Isolating the Brief Icon

There are two very significant comments to be made about the process that transpires when one looks at a briefly exposed (say, 50 msec) display of letters or digits and then reports on its content. First, indifferent to the total number of letters or digits presented, one can only correctly report the items in four or five locations of the display. Even so, one genuinely feels that he/she saw quite a bit more than he/she was able to recall. Second, the display appears to last much longer than it actually does.

One could inquire, on the one hand, about the fate of the information seen but not reported, and on the other hand, about the phenomenal duration of the exposure in relation to the actual duration of the exposure. In the former case, our inquiry is about information persistence, and in the latter, about phenomenal persistence. Both lines of inquiry have been followed in the attempt to divulge the character of the icon.

Information Persistence. The method of delayed partial-sampling is the major procedure for examining informational persistence. The method is to tachistoscopically display a number of items, usually letters or digits, that exceed the memory span, and to follow the display after a brief interval by an instruction to the observer to report a subset of the display. The subset that is called for is within the span of immediate memory, and the observer is ignorant of the subset he/she is to report until the instruction is given. The purpose of partial report is to circumvent the limitations imposed by short-term storage--the system held responsible for one's inability to report more than five items when one is attempting to report them all (cf. Sperling, 1960; Averbach and Coriell, 1961). The significant feature of this method is that the selective instruction, provided that it is given within milliseconds after the display, gives a measure of item availability superior to that obtained in the noninstructed case where the observer tries to report as many items as possible. Suppose that the display size is twelve items, and the subset size is three items. If the observer can report any subset specified immediately on termination of the display, then it is concluded that at the moment of comprehending the instruction, all twelve items were available to him (Sperling, 1960).

In the celebrated experiments of Sperling (1960) and Averbach and Coriell (1961), the large difference in estimated item availability between the immediately instructed (0 msec delay) and noninstructed conditions, in favor of the former, was taken to imply the existence of a large-capacity sensory memory. However, these experiments went further. By delaying the selective instruction, it was shown that partial report exceeded whole report by less and less; within a short interval of delay (less than a second), there was not a wit of difference between the two. This finding implied that the large-capacity memory had but a brief life, though the question of how brief has never been precisely solved. Indeed, the situation is even worse, as estimates range from fractions of a second to seconds.

In part, variations in these estimates of duration are due to the complex of processes operating in the procedure. One especially enterprising effort to achieve precision in the estimation of iconic informational persistence is due to Averbach and Coriell (1961), and to this insightful series of experiments we now turn, for as will be seen, it is illuminating in several significant respects.

Credit for full appreciation of the fact that performance in a delayed partial-sampling task is the result of two different types of performance on the part of the participant is due to Averbach and Coriell (1961). One type of performance is a nonselective readout, independent of the occurrence of the instruction; the other is a selective readout that occurs only subsequent to the decoding of the instruction. Nonselective readout is suggested by the fact that delayed partial-sampling performance never appears to approach zero; instead it asymptotes at the level of noninstructed or whole report. It is therefore assumed that the observed begins to enter material into a more permanent memory--the short-term store--as soon as possible, at least before the instruction cue is apprehended. On occurrence of the instruction cue, some of the designated material may have been processed already. Just how much depends on the size of the display and the overlap between preselected and cued items. Hence, following Averbach and Coriell's suggestion, it is hypothesized that performance in an iconic memory task is supported by two kinds of storage mechanisms: iconic storage and short-term storage. One is reminded of the parallel, though later, comments of Waugh and Norman (1965) with respect to the dual support (short-term storage and long-term storage) of performance in short-term memory tasks.

Considering this insight, it is asked how iconic storage time might be determined from the data obtained in an iconic memory task. Suppose a display of two rows of eight letters is exposed, and that this display is followed by one of the following: a bar that points to the location of one of the sixteen letters; a ring that surrounds the location of one of the sixteen letters, or a cross-hatched circle that spatially overlaps the location of one of the sixteen letters. In each instance the observer must attempt to report the signaled letter. Figure 7 captures an ideal observer's performances in each of these three situations. Essentially, Figure 7 is a composite of several figures from the Averbach and Coriell (1961) communication and a precis of several separate experiments. The bar-marker curve represents a reasonably typical performance under conditions of delayed partial sampling. The other two curves reveal an impairment in performance, with respect to bar-marker performance that we must attribute to some kind of perceptual impairment induced by the ring and the cross-hatched circle. This impairment will be referred to as masking, the ring and circle as masks. The signaled letter will be referred to as the target.

There are two simple but important lessons to be learned from Figure 7. First, it is apparent that in the delayed partial-sampling procedure, the cue for selection might also serve to impede identification. A second and related point is that the spatial relations between the target and mask influence the functions relating target identification to mask delay. There will be more to say about these observations later; for the present, note how Averbach and Coriell put them to use.



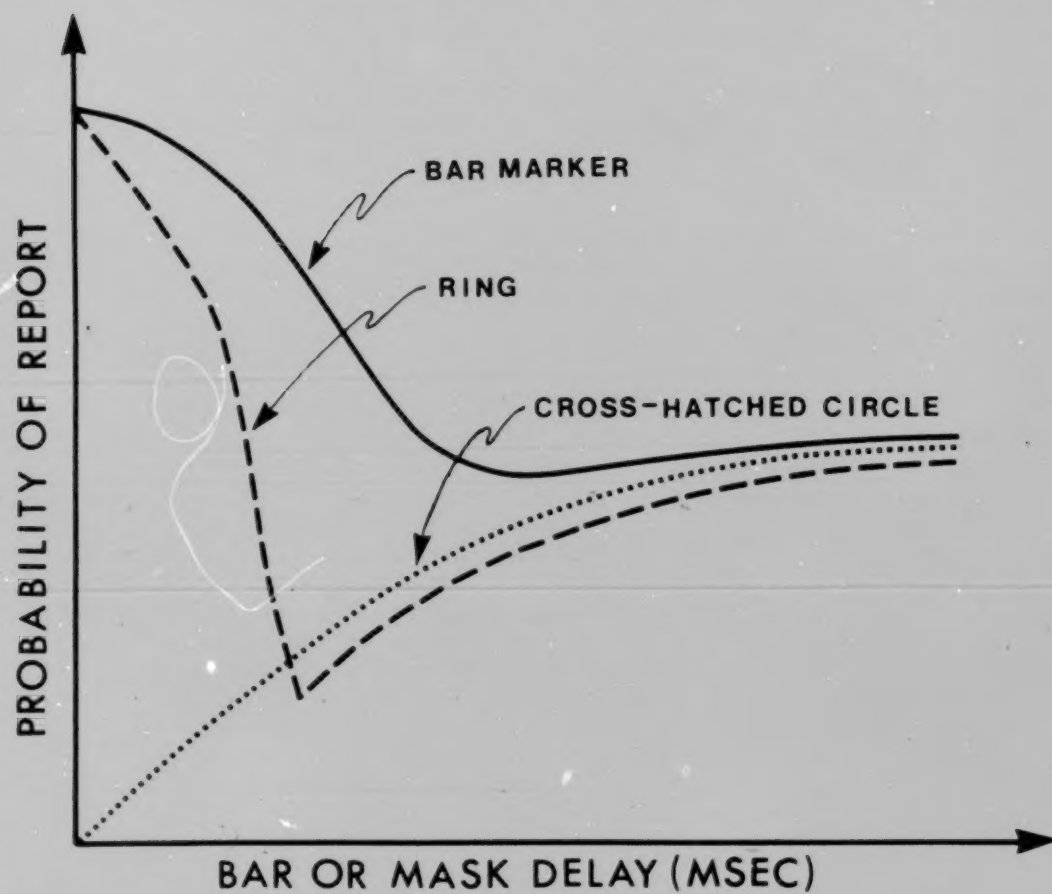


Figure 7: An idealization of the experimental results of Averbach and Coriell (1961).

To begin with, they interpreted the nonmonotonic function induced by the ring as saying that at very brief delays target letter and ring superimposed, but at longer delays the ring erased the target letter. In a similar vein, the cross-hatched circle was said to induce a monotonic function because, in the superimposition phase, the target is lost in the target-mask montage. On the assumption that a mask at some delay erases the iconic record of a letter, Averbach and Coriell conjectured that one ought to be able to estimate the temporal course that registers a selective instruction (say, a bar-marker), and translates the cued and iconically represented letter into a more unyielding representation. Their conjecture was realized in a paradigm in which a bar marker specifying a particular letter was presented simultaneously with a 2 x 8 display and followed at varying intervals by a ring surrounding the location of the specified item. The minimal delay permitting the unhampered report of the specified letter could be taken as the measure of the minimal time needed to register the cue and effect the letter's translation to short-term storage. Additionally, by this procedure, an estimate could be determined, at each ring delay, of the probability that an item was already in the short-term form. Averbach and Coriell were able by such means to compute the separate contributions of iconic and short-term storage to partial report, as a function of delay; and by such means, iconic storage was shown to persist for 250 msec.

This concludes the elementary discussion of the Averbach and Coriell (1961) experiments, and it remains to summarize the essential features. Apparently, the role of the selective instruction, or probe, in delayed partial-sampling is to direct the transformation of some items from the iconic form to the short-term form. However, as this transformation proceeds independently of the selective instruction, the role of that instruction must be to change the transformational process from one subset of items to another, that is, to switch attention. These two components of the task, processing the selective instruction and switching attention, can be shown to consume time (Erickson and Collins, 1969; Erickson and Colgate, 1971; Erickson and Hoffman, 1972)--time that is critical to accuracy of report given the rapid decay of the iconic representation. When the selective instruction or probe is visual--such as a bar-marker--performance is not only affected by the time to interpret the probe, but by the inability to determine precisely the probe's spatial relation to the matrix of letters (Townsend, 1974; Lowe, 1975). Delayed partial sampling performance can be enhanced (Lowe, 1975) by giving the probe a spatial framework, such as embedding it in an array of bar-markers of which it is the longest. It goes without saying that delayed partial-sampling is a bedeviled tool and we should not wonder at the anomalous data and interpretations that it occasionally spawns.

For example, there is a point of view that differences between partial and whole report are artifacts of scoring procedures (Holding, 1970; Dick, 1971). Usually, fewer items are to be reported in partial-report than in whole-report, so that the greater "output interference" (Tulving and Arbuckle, 1963) biases the proportionality measure against whole report. If one looks at short strings of letters in whole report and compares accuracy at each position in the string with accuracy at each position in strings of the same length given in partial report, then whole report is superior by this absolute measure (Dick, 1971). One could conclude as a consequence of

this analysis that there is no large capacity buffer. However, absolute performance in partial report must be lower than that in whole report (Bennet, 1971). In whole report, the observer chooses which items to relate--presumably those of high fidelity, while in partial report, the items reported are at the experimenter's whim and therefore likely to be of variable fidelity. In whole report, "recall" is not delayed by the necessity to process a probe and transfer attention. In sum, though output interference may favor partial report on the proportionality measure, partial report is detrimentally affected by variables to which whole report is uncaring. Of the two measures described, the proportionality and the absolute, the former is probably the more viable index. Let us conclude our discussion of delayed partial sampling at this point with the guarantee that the bedeviled nature of this technique will receive a further hearing in Section B (The Notion of the Icon as a Precategorical Memorandum).

Phenomenal Persistence. There are a wide variety of methods by which the duration of a person's conscious awareness<sup>2</sup> of a briefly exposed display can be estimated. Our sample, described below, is representative rather than exhaustive.

One method entails recycling a brief exposure of a figure and determining the slowest rate at which the figure is continuously visible; an offset to onset interval of 250 msec has been reported as the minimum (Haber and Standing, 1969). A second and more complicated method might be dubbed the "subtractive reaction time" technique (Briggs and Kinsbourne, 1972). In this method, the observer, on separate blocks of trials, presses a key in response to the onset and in response to the offset of a briefly exposed display. By subtracting the latency to onset from the latency to offset, an estimate of phenomenal duration is obtained. To date, phenomenal duration measured by the method of subtractive reaction time (Briggs and Kinsbourne, 1972; Erwin and Hershenon, 1974) falls considerably short of the estimates obtained by other means.

Ideally, as Efron (1973) remarks, one is seeking an instrument that would start and stop a clock at the beginning and end, respectively, of the phenomenal experience of a briefly exposed display. The method just described is an approximation of such an instrument. Another approximation follows from a procedure suggested originally by Sperling (1967). The observer adjusts two very brief sounds so that one is coincident with the onset, and one with the offset of a brief, but longer, visual exposure. The intersound interval is used to index the phenomenal experience. This method of indexing in a separate modality has provided estimates of phenomenal persistence ranging from approximately 130 msec (Efron, 1973) to 250 msec (Haber and Standing, 1970).

I must now remark upon two very significant features of phenomenal persistence. First and foremost, there is an often-observed inverse relation between the exposure duration and the phenomenal duration (Robinson, 1968; Haber and Standing, 1970). Indeed, we believe that within reasonable limits the relation can be stated formally as follows:

exposure duration + phenomenal persistence measured from exposure offset = a constant (Efron, 1973).

Efron has demonstrated, in elegant fashion, that this relation holds true in modalities other than vision, and thus appears to be a stock feature of the nervous system. Note that according to the above rule, when the exposure duration is equal to the constant, there is no phenomenal persistence. More generally, when exposures exceed the constant value, exposure durations and phenomenal durations are precisely equal, though of course, the phenomenal impression is temporally displaced and lags the optical event.

One might venture the hypothesis that the constant phenomenal duration identifies a significant constraint on the information-processing system, namely, for that system to complete its operations it is required that environmental occurrences be registered internally for a specified minimal period (cf. Haber, 1971; Efron, 1973).

I wish to remark upon the second feature that has to do with the contrast between phenomenal persistence as measured by one of the methods described above and subjective, perceived duration. The distinction being drawn is far from clear, but in an approximate way, we may consider subjective, perceived duration to be that which underlies an observer's behavior when, given two briefly exposed displays in succession, he simply comments that one display seemed to last longer than the other. There is evidence to suggest that variables which affect phenomenal persistence affect subjective perceived duration in the opposite fashion. I will have reason to consider the relation between these two measures of duration, but note, as a preliminary, that whereas phenomenal persistence may vary inversely with intensity, subjective perceived duration varies directly (Efron, 1974).

The Nonequivalence of Iconic Measures. Clearly, there is no standard way to investigate the "intuitive object" of a transient medium of literal storage. There are no handy rules for deciding which paradigm may lay claim to legitimacy; indeed, there is not even a rule of thumb. Nevertheless, it is easy to recognize why the different methods are likely to yield variant estimates and even dissimilar characterizations.

To begin with, the informational and phenomenal techniques are distinguishable by the kinds of processes necessary to their performance. Most obviously, the ensemble of processes manifest in the delayed partial-sampling procedure are not duplicated in the indices of phenomenal persistence. At the same time, the delicate task of establishing a criterion for when the icon has run its course is germane to phenomenal measures.

Because of its theoretical consequences, a more serious source of variance is created by the kinds of materials allowed in one paradigm but prohibited in another. Random dot patterns or dynamic visual noise (Uttal, 1975) can be used in phenomenal persistence paradigms, as described, and in paradigms that seek to measure iconic longevity by determining the maximal interval over which two fields may fuse into one (Eriksen and Collins, 1967; Pollack, 1973). Such optical arrangements, however, cannot be exploited easily for delayed partial-sampling. The theoretical quirk is that



conjectures about the form of the iconic representation are likely to be viable for the material one has displayed in a given procedure, but irrelevant to other kinds of material in other paradigms. Thus, the description of the icon as featural relations (such as oriented lines) is felicitous, in principle, for the form of storage supporting the delayed selection of letters or digits from an array of similar items, but it is nonsensical when applied to the representation permitting one to fuse an earlier-arriving dot stereogram with a later-arriving dot stereogram and thereby experience a form-in-depth (Ross and Hogben, 1974).

All these general considerations hint at the possible but opposite conclusions that there may be either a variety of sensory registers in vision, or one abstract description underlying all manifestations of persistence--informational or phenomenal.

### Nonvisible Visual Representation

The foregoing procedures seek to isolate a very special kind of visual representation that we have chosen to call "iconic." A most significant feature of this representation is that it is visible. There is another kind of visual representation to be considered that is most obviously not visible and to this class of representation we turn for the twofold purpose of delimiting its character and putting into sharper relief the form of the iconic representation.

An often expressed view is that the iconic representation of linguistic material undergoes a metamorphosis from a visual to a linguistically related form. One elegant expression of this sentiment suggests that the raw visual data are cast rapidly into a set of instructions for the speech articulators, for subsequent (and more leisurely) rehearsal and report (Sperling, 1967). In the more general case, however, in which the information to any eye is not about linguistic material but about surface layout, objects, and events, we would be surprised to find that the iconic conversion was solely a linguistic matter. Interestingly enough, it appears to be the case that even for linguistic--or at least alphanumeric--material, noniconic representation is not limited to a linguistically related form. What kind of evidence can be adduced for this claim? Putatively, it must come from experiments in which presentation of alphanumeric material is not excessively brief; exposure time ought to exceed the minimal, relatively invariant, duration we spoke of above. Further, the index of the representation should yield a persistence value significantly removed from the persistence values estimated for the icon. With these as crude criteria, consider the following experiments. When two letters are presented for comparison, the judgment latency is shorter if the comparison can be based on visual appearances (for example, A compared to A or B) than if the basis is nonvisual (for example, A compared to a or b). However, if the presentation is successive and the temporal separation is greater than several seconds, comparison latency in the two situations is virtually identical and at the magnitude of that based on nonvisual appearance (see Posner, 1969). The implication is of a gradual shift from an exploitation of a visual representation to an exploitation of a nominal one. An even more durable representation related to visual

appearance is exhibited in the circumstance where one must maintain visually presented and aurally presented letters during the course of verbal shadowing; after shadowing for tens of seconds, report of the visually presented is superior to that of the aurally presented (Kroll, Parks, Parkinson, Bieber, and Johnson, 1970).

Of course, we are justified in complaining that experiments such as those described suggest no more than a nonverbal source of maintenance for visually exposed letters. This source could be iconic, and previous experiments have simply underestimated its persistence. How might we, then, determine a difference between the two kinds of representation, assuming that there is one?

Fortunately, we have already made reference to the fact that whatever kind of representation is responsible for partial report superiority, that representation is maskable (cf. Averbach and Coriell, 1961; Sperling, 1963). Moreover, the apparent curtailing of persistence due to an after-coming mask holds for both informational and phenomenal estimates (see Haber and Standing, 1970). An extraordinarily good illustration of the relation between information persistence and masking is provided by Sharf and Lefton (1970): where whole-report of a 12-letter display was requested, delaying a mask for longer than 50 msec did not impede performance; in sharp contrast, partial report specified by a delayed selective instruction was impeded by a mask delayed for at least as long as 250 msec. Examination of partial-report performance as a function of the interval between onset of the selective instruction (a tone) and onset of the mask, revealed that performance did not increase with negative intervals (mask leading), but did with positive intervals (mask lagging). In the latter case, it is a notable observation that performance continued to increase up to separations between the instruction and the mask on the order of 300-400 msec. (Sharf and Lefton, 1970). We must adamantly conclude that there is a particular way in which an experience of a visual display may persist--dubbed iconic--that is highly sensitive to aftercoming, temporally proximate and discrete visual occurrences. We can now ask whether there are representations related to visual appearance that are not characterized by this sensitivity. In anticipation, the answer is yes.

Consider a situation much like the above, in which a man or woman looks at two successively presented, novel, and not easily verbalized patterns that are either identical or very similar. The task in this situation is to report whether the two patterns are the same. The participant can do remarkably well, as measured by either latency or accuracy, up to fairly long intervals, for example, 9 secs (Phillips and Baddely, 1971) and 20 secs (Cermak, 1971), even though a mask may intervene between presentations (Phillips and Baddeley, 1971; Mitchell, 1972). In sum, there is reason to argue from experiment as well as from intuition, that there is a form of representation for things seen that endures masking and that is nonverbal. These conclusions are collected together in an experiment by Scarborough (1972) that contrasts with those just described in its use of easily verbalized material.

People heard a sequence of symbols (letters or digits) and then saw a display of symbols that was followed by an instruction to report either the symbols heard or the symbols seen. The display of symbols was always followed by a mask. The concern was with the accuracy of comparison in report for the symbols seen (heard) in the context of symbols heard (seen), with the accuracy of report for symbols seen (heard) in the absence of symbols heard (seen). The significant result is that the delayed visual report for a masked display is virtually unimpaired by the presence of an auditory sequence. This is at odds with the point of view that alphanumeric material survives the icon in a form that is closer to audition than to vision. We could have supposed, quite reasonably, that auditory short-term memory is the posticonic medium of storage, inasmuch as iconic items are translated into "auditory" items by implicit speech (Sperling, 1963, 1967). If this were the case, then the concurrent retention of aurally presented material should have been especially damaging to visual performance.

These last remarks provide a point of entry, albeit devious, into the question of what limits performance in the report of items from a briefly exposed display; they will, in addition, lead us to a further distinction between "visible" and "nonvisible" visual representations.

A number of students of visual processing have proposed that the span of apprehension--more aptly, the actual number of alphanumeric items reported from a single glance--is owing to the limited capacity of short-term memory (Sperling, 1960; Estes and Taylor, 1964). Suppose for the time being that by "short-term memory" we intend "auditory short-term memory." It has been shown that the brief retention of items from a masked display is not seriously impeded by an auditory memory load (Scarborough, 1972). It can be added further, that the errors in report reveal little evidence of auditory confusion, in contrast to the evidence for visual confusion that is considerable (cf. Rudov, 1966; Wolford and Hollingsworth, 1974). From these two counts, it may be conjectured that an auditory-based, linguistically-related (Neisser, 1967) medium of storage is not the limiting factor in report from a single glance. Consequently, to suppose that the limiting factor is a representational medium that outlives the icon, necessitates a look to a nonvisible representation, like that examined above, as a possible candidate. In this regard, Sanders' (1968) observations are instructive. Suppose that, one at a time, some of the cells of a matrix display are filled. The number of marked locations an observer reports correctly is about three or four, and the memory responsible for this performance, unlike its verbal (auditory) counterpart, is not characterized by recency, is insensitive to presentation rate, is subject to visual confusion, and is affected by the set of alternatives (Sanders, 1968). It is not without interest that a similar capacity estimate for a reputedly noniconic, but visual representation, emerges from the investigations of Scarborough (1972) and Posner (1969), and that the number  $4 \pm 1$  identifies the number of items reported in their correct location after a single glimpse of a display of unconnected items (Henderson, 1972).

There are two conclusions to be drawn at this juncture: the limit on what can be reported at a single glance is not a memorial representation that arises from implicit speech, but a representation that is cognate with visual

appearance; this latter representation has a capacity that falls far short of the capacity of the icon.

We may now return our attention solely to the task of distinguishing the "visible" from the "invisible" representation; thus far, the former is characterized as maskable, of indefinitely large capacity and very brief, the latter as nonmaskable, of limited capacity, and temporally substantial. The distinction is further enhanced by the work of Phillips (1974). When two patterns are presented successively in a same-different task, accuracy and latency of judgment as a function of temporal separation is affected by the complexity of the patterns; performance declines more rapidly the higher the complexity. However, the decay function can be shown to consist of two components, only one of which is actually sensitive to complexity. There is an early component of high accuracy indifferent to complexity that is nonoperative if the patterns in a pair are spatially displaced; by displacing the lagging pattern, accuracy falls to a level determined by a second long-lasting component that, though insensitive to movement, is sensitive to complexity. In the domain of the short-lived, high accuracy component, a difference in successive patterns induces in the observer an experience of "seeing the shape change;" in the domain of the durable, less accurate component, the same difference is experienced as "knowing that the patterns changed," rather than as seeing the change (Phillips, 1974). Interestingly, of the two components, only the position-sensitive component is markedly maskable. We see, in short, that there is a further and potentially significant distinction to be drawn between the so-called visible and nonvisible visual representations: the former is tied to a fixed position (defined probably in the coordinates of the retina), the latter is not.

Although this inventory of contrasts between visible and nonvisible visual representations is most likely incomplete, it is evident that there is a worthwhile distinction to be made. Let us therefore distinguish two modes of persistence for brief optical stimulation: iconic and schematic.

#### The Notion of the Icon as a Precategorical Memorandum

We have remarked several times that the iconic representation is precategorical in the nontrivial sense that it is a literal description and not a semantic or symbolic description of the structured light at the eyes. This is no more than an in theory claim and it is now incumbent upon us to see whether we can provide the proof.

The participants in Sperling's (1960) original experiments were asked to partial report by row. Obviously they found this instruction relatively easy to comply with and we might conjecture that, in part, it was because spatial location is a structural characteristic of the icon. Presumably, if they had been asked to report according to some criterion that was not structurally referenced by icon, then their performance in partial report would have been less than adequate. Sperling (1960) presented his observers with a brief display of a mix of letters and digits preceded by an instruction to report one or the other category. The outcome was that the participants in the experiment performed abysmally--their partial report scores were no better



than their whole report scores.

One interpretation of this result is that while spatial distinctions are made at the level of the icon, category distinctions are not. Thus, iconic storage can be accessed by spatial location, but not by category. In terms of the flow of information depicted in Figure 1, to determine that an item is a letter or a digit requires the services of the pattern recognizer, and consequently the status of an item as a letter or a digit cannot be a basis for selection prior to pattern recognition, but only subsequent to pattern recognition.

The contrast between selection by spatial location and selection by the letter/digit distinction is, of course, the contrast drawn by Broadbent (1971), and remarked upon earlier, between filtering and pigeon-holing. In filtering, each of the items registered iconically is examined for the presence of a specified feature; where the feature is absent, no further examination is needed. In contrast, pigeon-holing is considerably more complex, since evidence from each iconically registered item must be considered by the limited-capacity categorizer to produce one or more of a prescribed number of category states. In sum, in Broadbent's (1971) analysis, rejection of irrelevant material is laborious and, we should suppose, more time consuming under conditions of pigeon-holing than under conditions of filtering. We can imagine, without difficulty, therefore, how selection by pigeon-holing might exceed the useful life of the icon.

There are other observations that apparently support the general conclusion that filtering is more efficient than pigeon-holing. In two notable experiments, von Wright (1968, 1970) demonstrated that delayed partial report by one of the following criteria--location, color, size and brightness--was superior to whole report. Further, von Wright (1968, 1970) found no evidence for efficient selection using letter orientation or category (letters vs. digits, consonants vs. vowels). There have been other demonstrations of the efficiency of color as a selection criterion (Clark, 1969; Turvey, 1972), and in addition, we may note that selection by shape (Turvey and Kravetz, 1970) and selection by direction of movement (Triesman, Russel and Green, 1974) leads to a significant difference in partial and whole report, in favor of the former.

We draw attention to the fact that, in the experiments just noted, the degree of difference between partial and whole report is used to index the status of a dimension as a selection criterion and, in turn, its status as a structural property of the icon. Inasmuch as semantic class proves to be an unsuccessful basis for selection, we conclude that semantic class is not a characteristic of iconic representation. Further, if it is observed that partial report by criterion x is significantly more superior to whole report than partial report by criterion y, then it might be concluded that dimension x is a more salient iconic dimension. Unfortunately, in the following, it will become apparent that conclusions of this kind do not necessarily follow from the partial report-whole report difference.

Consider two experiments that required partial report from eight-item displays consisting of four red items and four black items, four of which

were letters and four of which were digits (Dick, 1969, 1970). In both experiments, selection by semantic class was better than selection by row, which in turn was better than selection by color. This ordering, of course, is contrary to that evident in the results reported above. However, by noting item and location uncertainty, we may soften, if not resolve, the contradiction. The mixture of letters and digits is high in location uncertainty but low in item uncertainty. The letters and digits were drawn from sets of eight, and in consequence, the selective instruction halved the number of possible responses. Given that the participants in these experiments (Dick, 1969, 1970) were confronted with this small set of responses and were permitted to free recall (that is, report items indifferent to location), then we may interpret the greater success of selection by class as due, in very large part, to guessing.

What of the contrast of report by row and report by color? Primarily the difference between the two conditions is this: in report by row the desired items occupied proximate and connected locations; in report by color, the desired items were scattered haphazardly about the display. Report by row is low in location uncertainty but high in item uncertainty; report by color is high in both uncertainties. If we assume that item uncertainty is less detrimental to performance than location uncertainty, then we can appreciate how the ordering of selection criteria observed by Dick (1969, 1970) mimics the ordering of uncertainties (cf. Bennett, 1971). The lesson to be learned is that in the experiments reported in this section, selection criteria are confounded with the uncertainties of items and of locations. For any given selection criterion, the form of this confounding determines the partial report-whole report difference.

In this respect, an experiment by Fryklund (Bennett, 1971; Fryklund, 1975) is especially illuminating. The criterion for selection was color--one had to report the identities of five red letters in a briefly exposed 5 by 5 matrix--and this criterion was held constant across conditions. Of interest was the partial report performance as a function of the spatial arrangement of the red letters and the similarity of the target and the background items, that is, the items filling the remaining cells of the matrix. Partial report proved to be significantly affected by both variables. Of the two, target-background similarity was the more important constraint on performance, with the spatial arrangement serving to control the degree of interference from similar background items.

The implication, in short, is that we are in no position to adjudicate among selection criteria in terms of their efficiency. For it is obviously the case that in all of the aforementioned experiments from which we wish to draw conclusions about selection efficiency and the structure of the icon, the spatial arrangement of the targets and their similarity to the foils were uncontrolled (but manifestly operative) variables. Broadbent (1971) believes that filtering will be superior to pigeon-holing in most instances. And though this is a conclusion which we may abide, it is an unwelcome fact that we cannot with certainty claim this conclusion in the domain of iconic storage. The data available to date do not permit us this luxury; partial report-whole report differences are not unbiased sources of selection efficiency.

Maintaining this negative attitude, we can note that the compatability of instruction to response is a significant factor in the delayed partial-sampling procedure. A high tone for the top row of a matrix and a low tone for the bottom row, or a bar marker spatially adjacent to the desired item, represent conditions of high compatability when contrasted with one tone for letters and another for digits. In the latter case, time to interpret the instruction will be considerably longer than in the former cases; and as we can readily understand, partial report performance is sensitive to the time taken to encode the instruction probe (for example, Eriksen and Collins, 1969).

This argument suggest that, on the basis of partial report-whole report differences, we cannot calim that semantic distinctions are iconically absent.

In conclusion, the promisory note relating to the proof of iconic precategoricity must remain unfulfilled.

#### On the Malleability and Docility of the Icon

We turn now to an issue that bears closely on the question of categoricity, namely, whether the icon is malleable and docile. Behind the claim that the icon is precategorical, is the idea that the icon is indifferent to and largely separate from the structures and functions that constitute the observer's knowledge about the world. Accordingly, the icon is said to be passive, that is, its maintenance is not parasitic on cognitive resources, and that whatever may be its nature, it is not fashioned by mechanisms affiliated with and modulated by the longer-term memory. This contrasts with short-term storage that is said to be a rehearsable categorical memorandum and therefore most intimately connected to one's long-term model of the world. Indeed, it is useful to conceive of memory in the short-term as the temporary activation and organization of permanent memorial representations (Norman, 1968) and this conception provides a departure point for our consideration of iconic flexibility.

Some years ago, Hebb (1949) proposed that memory over the short-term differed from memory over the long-term, in that the former is a dynamic and transient reverbatory trace initiated by stimulation, while the latter is enduring and structural. If reverberation continues unabated for some goodly period, then the information dynamically represented becomes instantiated structurally.

Suppose that a supraspan list of digits is presented for immediate recall at a rate (say, one digit per second) that permits the viewer or listener to attend to each digit. In Hebb's view, the retention of the list is by virtue of a specific dynamic relation among the permanent structural representations of the digits. Consider what ought to transpire when immediately following this list's recall, a second list is presented that is a different permutation of the same set of digits. According to the theory, the dynamic activity supporting the retention of the first list is annihilated and a new pattern of activity, one tailored to the new

permutation of the same set of digits, is established. In other words, the presentation of the new permutation for recall eliminates the dynamic record of the previous permutation. By this reasoning we would expect, therefore, that in a series of such immediate memory tests, if a particular sequence of digits is repeated, say, every other or every third test, recall accuracy for the repeated sequence would be no different from that of nonrepeated sequences. Both Hebb (1961) and Melton (1963) presented results contrary to this expectation and concluded that any dynamic trace necessarily produces a permanent structural representation.

Note, however, that in a reasonably representative information-processing scheme, such as depicted in Figure 1, we distinguish two variations of relatively brief memory: one that is prior to attention (iconic), and one that is subsequent to attention (short-term), where attention for these purposes is virtually synonymous with categorization (cf. Broadbent, 1971). According to this distinction, it might be claimed that the particular "dynamic trace" assayed by Hebb and Melton was that subsequent to attention. If so, it may be inquired whether the hypothesized activity supporting the retention of material anticipating attention similarly induces a structural change.

Ideally, in order to make this inquiry, a delayed partial-sampling procedure is needed in which a particular set of items is repeatedly presented to an observer--in a situation in which the set is never selected for identification and report, though other nonrepeated sets of items, are so selected. In an experiment that approximates this ideal (Standing and DaPolito, 1968), a string of letters was repeated in successive briefly exposed displays, although the string's position in a display was varied. Partial report was always of another row containing a nonrepeated string. A subsequent test revealed that several repetitions did not enhance recall of the critical items. Unfortunately the repetitions were few in number, and in consequence, this single result is not very convincing. This inquiry can be pursued by looking at experiments that less closely approximate the ideal.

Instead of repeating a position-varying sequence of letters in successive displays, the experimenter could repeat an entire display and arrange matters such that delayed partial report of the same row on consecutive repetitions is rare; in the first repetition the middle row is cued, in the second, the top row, and so on. In principle, this means that with each repetition the same description (whatever form it might take) is entered into iconic storage, but that varying and often inaccurate descriptions (given delays greater than zero) are entered into the short-term store. Even if reports were perfectly accurate, by alternating repetitions with presentations of novel displays, the number of row reports intervening between repetitions of the same row report becomes excessive [in view of Melton's (1963) demonstration that Hebb's (1961) effect disappears with increasing distance between repetitions]. In sum, under these conditions the source of a repetition effect, if any, would have to be the icon.

The outcome of such experiments is that when observers are ignorant of the fact of repetition, partial report of a repeated display is not superior to partial-report of nonrepeated displays. Both Turvey (1967) and Merluzzi



and Johnson (1974: Exp. 3) found no enhancement with 54 repetitions; Glucksberg and Balagura (1965) failed to detect an effect with hundreds of repetitions.

What does it mean for partial report to be indifferent to repetition? First, it implies that the iconic state per se does not have any long-term consequences. If we adhere to the assumption that the icon is precategorical and take into account the results of Hebb and Melton, then the preceding sentence is tantamount to saying that categorization is a necessary condition for effecting a permanent change in long-term memory (cf. Broadbent, 1971). (We will have more to say about this later.) Second, it suggests that the icon per se is not docile. If this is true it is not surprising, given what we intuit the icon to be (a sensory buffer with a rapid turnover of information). Furthermore, it does not necessarily distinguish the icon from short-term storage. All we know is that experiments using a short-term memory procedure yield repetition effects in the form of more permanent memory; we do not know that repetition affects the hypothesized short-term store, and by all accounts it should not. Let us pursue this second implication a little further.

Recall the argument that partial report is supported by two stores: iconic and short-term (cf. Averbach and Coriell, 1961). In the temporal range of delayed partial sampling experiments, the contribution of the iconic drops off sharply, while the contribution of the short-term is relatively constant. As we have seen, the partial-report whole-report difference may be taken as an index of the state of the icon. It proves to be the case that if whole-report trials are mixed with partial-report trials, partial report performance is enhanced by repetition (Besner, Keating, Coke, and Maddigan, 1974; Merluzzi and Johnson, 1974). However, the source of this enhancement is not the icon, but the increased contribution of the short-term component that, after Hebb (1961), can be said to reflect the increasing structural representation, that is, permanent memory. Here is the evidence: the whole report of a repeated display is superior to that of nonrepeated displays; significantly, though, repetition affects neither the magnitude of the partial report-whole report difference (Besner et al., 1974; Merluzzi and Johnson, 1974), nor the rate of decline in that difference (Merluzzi and Johnson, 1974). The conclusion is inescapable: repetition is not a variable to which an informational persistence measure of iconic storage is sensitive.

If the icon does not learn, it is at least malleable in the sense that its parameters can be manipulated. Parameter changes ought to follow from changes in display luminance and display duration. The relations between these energetic variables and iconic character, both informational and phenomenal, have been treated fully elsewhere (Dick, 1974), and we will touch upon them only by way of summary. In general, exposure duration is immaterial: beyond some minimal value, exposure duration does not affect the personality of the icon (cf. Haber and Standing, 1970; Sperling, 1960; Efron, 1974). Luminance is a much more slippery case, though we can reach a fair compromise; luminance affects performance level and persistence, but not necessarily the rate of decline (Dick, 1974).

To contrast, there is a nonenergetic approach to the icon's malleability, one that is especially tailored to the interpretation of the icon as a feature description.

Populations of neural systems that are selectively sensitive to featural relations--such as the orientation, size and movement of a bar of light--can be modulated by cross adaptation. Essentially, this technique demonstrates a loss of sensitivity to one pattern given preceding exposure(s) to another. This measure of change in sensitivity contrasts adaptation with the "after-effect" that is concerned with the degree to which perception is reversed, though we intuit that the same structures underly both. Dick (1972, 1974) has hypothesized a conceptual intimacy among after-effect, selective adaptation, and iconic storage, and there is evidence to bolster his hunch; however, the nature of the relations among these concepts remains opaque.

When a grating is viewed for some period of time, the thresholds for the same and similar gratings are raised, but thresholds for gratings differing in orientation and size are virtually unchanged (for example, Blakemore and Campbell, 1969). Such effects may be interpreted as due to a reduced responsiveness in the population of neural systems responsible for the detection and signaling of these features. If by the phenomenal persistence of the icon we mean, in part, a persistence in the signaling of features, then a reduction in the responsiveness of feature-signaling systems ought to be accompanied by a reduction in persistence. This was the logic behind a pioneering experiment by Meyer, Lawson, and Cohen (1975). They recycled a brief exposure of a grating (horizontal or vertical) and determined the slowest rate at which the grating was continuously visible. Measured in this fashion, persistence decreased following adaptation to a grating of the same orientation, and increased subsequent to adaptation to a grating oriented orthogonally; a result that held whether adaptation took place on the eye receiving the test grating or on the eye opposite. The implication is that the neuroanatomic origins of this effect are fairly central.

As hinged at the outset of this section, a major implication of the conjectured schism between the iconic store and the cognitive apparatus is that the processing limitations that apply to the system as a whole do not apply to the icon. The icon is the data base; and it is the processes operating on that data base that compete for the available resources. Following Norman and Bobrow (1975), a process may be defined as a set of programs that are conducted with common purpose and for which various resources--such as different kinds of memory capacity and communication channels--are appropriated as a unit. With regard to our current concerns, a process may be limited by the data provided iconically and/or by resource availability (Norman and Bobrow, 1975).

Suppose, however, that contrary to the orthodoxy expressed above, the icon, as a medium of storage, is sensitive to resource limitations. That is to say, suppose that maintaining the icon is a process that is parasitic upon the availability of resources. Then we would expect that a task performed concurrently with maintaining the icon should affect, detrimentally, the persistence of the icon. Short-term retention drops off precipitously when it is in competition with another process (cf. Turvey and Weeks, 1975);

perhaps the decay rate of the icon will be similarly affected.

There is a relatively simple strategy to approach this issue. Give a subject several items to be read (or listened to) and retained for several seconds. In between presentation of these items and their recall, briefly show the subject a display followed soon by an indicator to report part of the display. In short, fill the retention period of a short-term memory-distractor task with a delayed partial-sampling task as the distractor, and see how performance on either is affected (Doost and Turvey, 1971; Chow and Murdock, 1975). The primary results of experiments using a strategy of this kind are: the accuracy of partial report at each delay of the delayed partial-sampling task is depreciated by the demand to retain material concurrently in the short-term form (Chow and Murdock, 1975), but the rate of decline in accuracy manifests an indifference to the reduced availability in resources (Doost and Turvey, 1971; Chow and Murdock, 1975). It seems, therefore, that the resource requirements of retaining, say, several letters for several seconds, affects the short-term storage component of partial report, but not the iconic storage component. In sum, the informational capacity of the icon and its persistence are unrelated to resource limitations. What appears to be affected is the process responsible for transforming material in the iconic form to the short-term form (Chow and Murdock, 1975, 1976).

The inquiry into iconic malleability now passes, with some difficulty, to two anomalous but provocative discoveries--prefaced by the following observation. There is evidence that immediately subsequent to its brief exposure, a familiar arrangement of letters consumes more of the limited capacity than an unfamiliar arrangement, but that the relation reverses within a second (Mewhort, 1972). One partial reading of this evidence is that the more familiar the material, the more rapidly it is mapped from the iconic into the short-term representation (cf. Phillips, 1971), and in consequence, the capacity of short-term storage is filled that much sooner.

If iconic storage is truly an interface, as some have suggested (for example, Turvey, 1973), then it should be coupled operationally to the systems that it interfaces. As an illustration, when short-term storage is filled, the icon shuts off. A notion such as this is apropos the anomalous finding that phenomenal persistence, as measured by the technique of subtractive reaction time, is an inverse function of the amount of redundancy (that is, level of information) in a display of letters (Erwin and Hershenov, 1974; Erwin, 1976). This relation holds only when the items are to be reported; in the absence of a report requirement, measured persistence is virtually identical for letter arrays that vary in their order of approximation to English. One is tempted to conclude that two modes of storage support phenomenal persistence: one that is insensitive to the information displayed, and one that is a function of the processes operating on that information. In that the two modes are phenomenally indistinguishable, it is the latter that sets the upper limit on persistence. By implication, the persistence of the icon is yoked to the processes conducted upon it.

There is, however, a different interpretation of this result that is more consonant with the orthodox view of the icon and that is owing to the second of the two aforementioned, anomalous, but provocative discoveries. Consider a situation in which participants must judge which is the longer in duration of two 30 msec displays presented one second apart, where the displays vary in the familiarity of their content (Avant and Lyman, 1975; Avant, Lyman and Antes, 1975). Experiments using variants of this situation reveal that subjective duration is affected by familiarity, the less familiar material perceived as the more durable. More surprisingly, this conclusion holds (and more markedly so) even when the brief exposures are sandwiched between masks so that actual identification of their content is at chance.

In the view of those responsible for this observation, where the contact between stimulation and the long-term model is incomplete, the rudimentary results of the early processing are communicated to a central monitor as temporal extents. The shorter apparent duration of familiar material reflects the greater automaticity, or directness, of processing such material. We are told, in short, that subjective duration is an index, and a sensitive index at that, of processing (cf. Ornstein, 1969; Thomas and Weaver, 1975); it is reflective of what has been done and it serves to distinguish even between different orders of processing that remain incomplete.

Let us now return to the subtractive reaction-time technique and ask: What does it measure? A hasty and earlier reply is that it measures iconic persistence, but we cannot be certain any longer of this answer; indeed, it is just as plausible that the technique indexes subjective duration. As remarked, Erwin's results with the subtractive reaction-time technique could be interpreted as revealing a storage medium that fluctuates with the amount of information to be processed. If the technique measures subjective duration, then the more apt conclusion is that reaction-time differences as a function of information-level differences are epiphenomenal concomitants of differences in processing qua processing and are quite blind to the longevity of the hypothesized iconic store. A less awkward phrase conveys this point of view: processing effort varies with information level, but not iconic persistence. The technique of subtractive reaction-time reflects, indirectly, the former.

Let me conclude this section with an earnest reflection on the aforementioned and related claims that the icon per se does not have any long-term consequences, and that categorization is necessary to the achievement of such. In a succession-of-stages model of the kind we are considering, it is not unreasonable to venture that particular memorial or perceptual consequences depend on information reaching a certain stage of processing. It is guessed that prerequisite to achieving a long-term memorial representation, information must be coded into short-term storage, and supposedly, information enters that store when attended. Against this notion, however, is the experiment of Cohen and Johansson (1967) which shows that the Hebb repetition effect does not occur when a list is fully attended to on each repetition, but is never repeated back. Although this experiment was conducted in the auditory mode, the result is significant to our current visual concerns. In terms of the stages model, repeatedly entering and



maintaining a list of things in short-term storage is not sufficient to determine a long-term representation.

Perplexities of this order have led some memory theorists to question the discrete stages model and to propose in its place the idea that there is a continuum of memories or persistencies, where persistence depends critically on the qualitative nature of the encoding operation performed (cf. Craik and Tulving, 1975). Aside from suggesting that we cannot talk meaningfully about the stage that information has to reach in order to have a chance of long-term storage, this critique of the stages model suggests that there cannot be an iconic representation. There are as many persisting memories, from brief to long, as there are different ways of interacting with a briefly exposed display. Curiously, the argument implies that what we have called iconic memory and schematic memory cannot mediate encoding but, rather, are the consequences of it.

#### Two Field Theory: An Introduction

For purposes of organization, most, if not all, of the discussion on the icon can be subsumed under the rubric of One Field Theory. The major questions and issues have focused on the immediate visual consequences of a single brief exposure; to put it precisely, our interest has been with processing two-dimensional displays in a single glance at a stationary point of observation. Let us now proceed to consider the perceptual consequences of two brief and independent exposures--two visual fields--temporally concatenated in various ways and presented at the same location and to the same point of observation. An exegesis of this, the two-field case, will borrow from One Field Theory; it will also reciprocate. Though we may have reason to introduce new concepts to account for two-field interactions, these concepts will serve to enrich our understanding of the single-field instance.

The most common phenomenon arising from the temporal concatenation of two visual fields is, of course, masking, a term that refers to the fact that the observer fails to see one of the fields as accurately as he would if it were presented alone. Most generally, Two Field Theory (cf. Pollack, 1973) considers only one temporal concatenate, the case where one field is presented after the other has ended. Where the first field impedes the seeing of the second, the term forward masking is used, where the second impedes the seeing of the first, the phenomenon is labeled backward masking. However, in view of the fact that successive fields can be partially or totally overlapping in time and that relations among the onsets and the offsets of the fields are most crucial to the effects obtained (Breitmeyer and Ganz, 1976), a precise and expanded differentiation of the temporal concatenation of two visual fields is needed. The following classification is recommended; it is a modification of a system that has proven useful to the analysis of temporal relations in movement (Golani, in press).

Let  $F_1$  and  $F_2$  be two visual fields, more precisely, two visual displays with the understanding that, in general,  $F_1$  is the target display, that is, it contains something the observer is trying to report, and  $F_2$  is the mask display, that is, it is meant to impair target report.

(1) If  $F_1$  and  $F_2$  follow each other without temporal overlap, the relationships will be referred to by the suffix "vene."

(2) If within the duration of  $F_1$ ,  $F_2$  occurs, that is,  $F_2$  either starts together or later than the start of  $F_1$ , and ends together with or earlier than the end of  $F_2$ , the relationship will be referred to by the suffix "dure."

(3) If  $F_1$  starts before  $F_2$  and ends after  $F_2$  started, but before or together with the end of  $F_2$ , the relationship is referred to by the suffix "vade."

(4) If  $F_1$  starts together with or after the start of  $F_2$  but before  $F_2$  ended, and ends after  $F_2$  ended, the relationship is referred to by the suffix "cede."

(5) If  $F_2$  occurs only during the middle portion of  $F_1$ , that is,  $F_1$  starts before  $F_2$  started and ends before  $F_2$  ends, the relationship is referred to by the suffix "case."

Through the use of appropriate prefixes we can distinguish among variants of these five cases: pre-(before) and super-(after); in-(going in) and ex-(going out); pri-(prior) and post-(later); con-(together); ent-(within); and en-(around). The combinations are collected together in Figure 8 and depicted in Figure 9.

In the terminology we have just developed, the  $F_1/F_2$  relation for forward masking is designated as "supervene" and that for backward masking is designated as "prevene." These are the most commonly investigated relations, as remarked, although others have been examined. "Entdure" is a case in point. When two fields are concatenated in entdure fashion (see Figures 8 and 9), it is possible to manipulate the intensity differential such that the considerably briefer  $F_1$  target exposure cannot be seen against the steady  $F_2$  "background" mask exposure. However, if the background field is terminated shortly after the offset of the target field, the target becomes visible (Standing and Dodwell, 1972; Turvey, Michaels, and Kewley-Port, 1974).

The point is that a complete Two Field Theory will have to speak eventually to each of the above temporal concatenations of  $F_1$  and  $F_2$ . At all events, let us note that given suitable structural relations between  $F_1$  and  $F_2$ , both supervene and prevene conditions yield masking when two fields are presented to the same eye (monoptic viewing) or to opposite eyes (dichoptic viewing). The advantage of contrasting these two ways of presenting the fields is that in the monoptic case, masking can originate in either relatively peripheral or relatively central neuroanatomic locations, while in the dichoptic case masking must be, necessarily, of relatively central origins.

|                                                                              | F <sub>1</sub> ends just<br>before F <sub>2</sub><br>starts | F <sub>1</sub> ends after<br>F <sub>2</sub> started but<br>before F <sub>2</sub> ended | F <sub>1</sub> ends<br>together<br>with F <sub>2</sub> | F <sub>1</sub> ends<br>after<br>F <sub>2</sub> ended |
|------------------------------------------------------------------------------|-------------------------------------------------------------|----------------------------------------------------------------------------------------|--------------------------------------------------------|------------------------------------------------------|
| F <sub>1</sub> starts before<br>F <sub>2</sub> started                       | Prevene                                                     | Invade                                                                                 | Convade                                                | Encase                                               |
| F <sub>1</sub> starts together<br>with F <sub>2</sub>                        |                                                             | Pridure                                                                                | Condure                                                | Concede                                              |
| F <sub>1</sub> starts after<br>F <sub>2</sub> started but before<br>it ended |                                                             | Entdure                                                                                | Postdure                                               | Excede                                               |
| F <sub>1</sub> starts immediately<br>after F <sub>2</sub> ended              |                                                             |                                                                                        |                                                        | Supervene                                            |

Figure 9: Two field concatenates (After Golani, in press).

Inspection of Figure 9 suggests that there are a great number of potentially significant temporal parameters. Of these, only a few have received serious study--primarily, those of special relevance to the prevene and supervene conditions: the durations of the two fields; the interval elapsing between the offset of the leading field and the onset of the lagging field (the interfield interval); and the interval elapsing between the onset of the lagging and the interval elapsing between the onset of the leading field and the onset of the lagging field (the field onset asynchrony).

With these conventions at our disposal we can identify two major functions determining perception in the prevene and supervene conditions. One function relates the energy (duration x intensity) of either the leading or lagging field to the interfield interval (IFI). In longhand, this function says that field energy multiplied by the minimal interfield interval (IFI min) needed to evade the perceptual impairment induced by the other field, is a constant. In shorthand,  $F \text{ energy} \times \text{IFI min} = \alpha$ . If we write this rule in the form,  $f:F \rightarrow \alpha F^{-1}$ , (where  $f$  is simply a function notation), we can appreciate that it is a special case of the exponential function  $y = ax^b$ , that is, the power function championed by S. S. Stevens (1970) as the exemplary psychophysical function. The significance of a power function is that it preserves ratios; in the case of  $f(F)$  above, constant  $F$  energy ratios yield constant minimal interfield interval ratios.

But why a power function? Inasmuch as vision must cope with enormous ranges of energy--often in excess of  $10^{12}$ --then a low exponent on a power function provides a compressor action. As Stevens (1970) envisaged it, the power function is nature's way of providing sufficient nonlinearity to effect

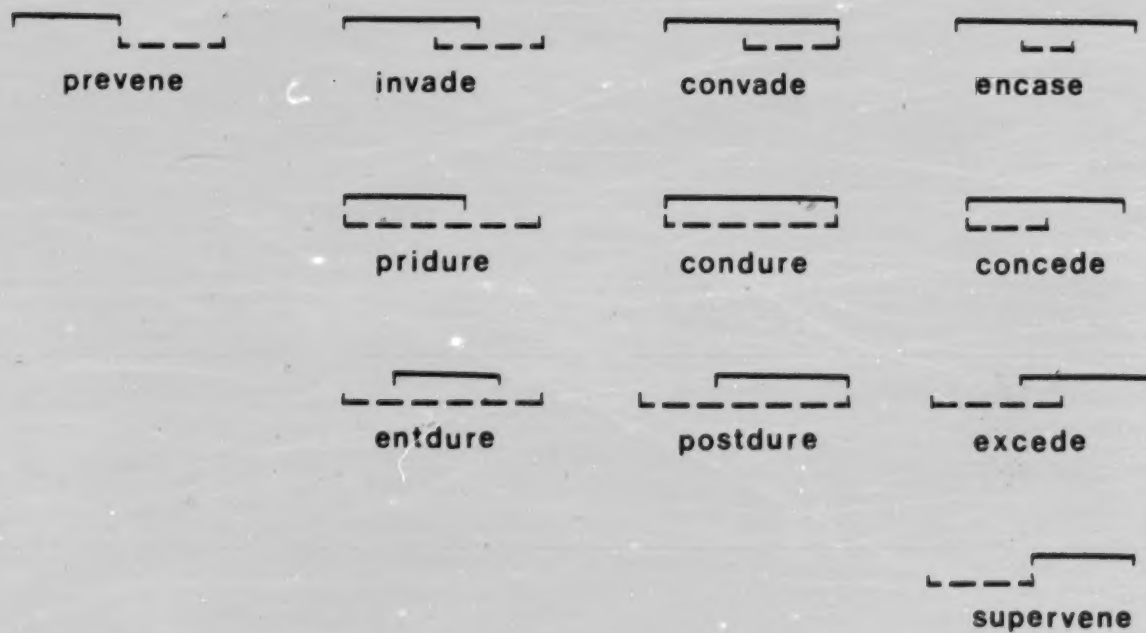


Figure 9: A depiction of two-field concatenates (after Golani, in press).



a match between the far-flung energy variation in the world and the processing capabilities of the nervous system. Intuitively, the transformation responsible for bending the sensory function by a ratio-preserving function is imposed at the eye, the first possible processing stage. However, more central sites are not excluded (Stevens, 1971). It is of interest that the masking power function has been obtained dichoptically (Kinsbourne and Warrington, 1962b), as well as monoptically (Kinsbourne and Warrington, 1962a; Turvey, 1973; Novik, 1974). It is also significant to note that the exponent of  $f(F)$  may vary from -1 (see Walsh and Till, 1975), and while the source of this variation remains to be explained, it lends credence to the claim that the two-field function in question relates more closely to the law Stevens had in mind, rather than the one Weber had in mind.

The second major function relates the duration of the leading target field to the interfield interval. Essentially, it states that for a given perceptual effect, the duration of the leading field and the interfield interval are complementary:  $F_1 \text{ duration} + \text{IFI} = \alpha \text{ constant}$ . Quite simply, the second function identifies field onset asynchrony as the significant variable rather than field duration, field energy, or interfield interval (Sperling, 1971; Turvey, 1973). Importantly, the function is obtained in both monoptic and dichoptic viewing (Turvey, 1973), and one should recognize that this function is the same as that described by Efron (1973) for phenomenal persistence (see Section B, The Notion of the Icon as a Precategorical Memorandum). The description of the two functions continues, for they bear significantly on questions of visual processing at a glance.

The power function is evident only where a necessary condition for masking is that the impeding field be of equivalent or greater energy than the target field. The function holds for both forward and backward masking, that is, for both the supervene and prevene conditions, although the constant is greater for the forward case (Kinsbourne and Warrington, 1962b; Turvey, 1973). For either condition, beyond some not-too-considerable energy value, the target field is immune to masking. In sum, in the domain of the power function, comparative energies of the fields is a more significant variable than the order of the fields, although, *ceteris paribus*, a leading field is more dominant than a lagging field.

The preceding summary is put into perspective when one considers masking in the domain of the additive function, for the function itself holds only for the prevene condition, that is, backward masking. Given prevene conditions that reveal the rule, one can compare the degree of masking in the prevene condition with the degree of masking in the comparable supervene condition. Put simply, the comparison reveals that a mask field impairs perception of a leading target to a greater degree and over a greater temporal range, than it impairs the perception of a lagging target, that is, backward masking is more pronounced than forward masking (Turvey, 1973; Uttal, 1975). Further, the energy of the following field need not be equal to or greater than that of the leading target field in order for perceptual impairment to occur (Turvey, 1973; Walsh and Till, 1975). In sum, in the domain of the additive function, the order of the fields is a more significant variable than the comparative energies of the fields.

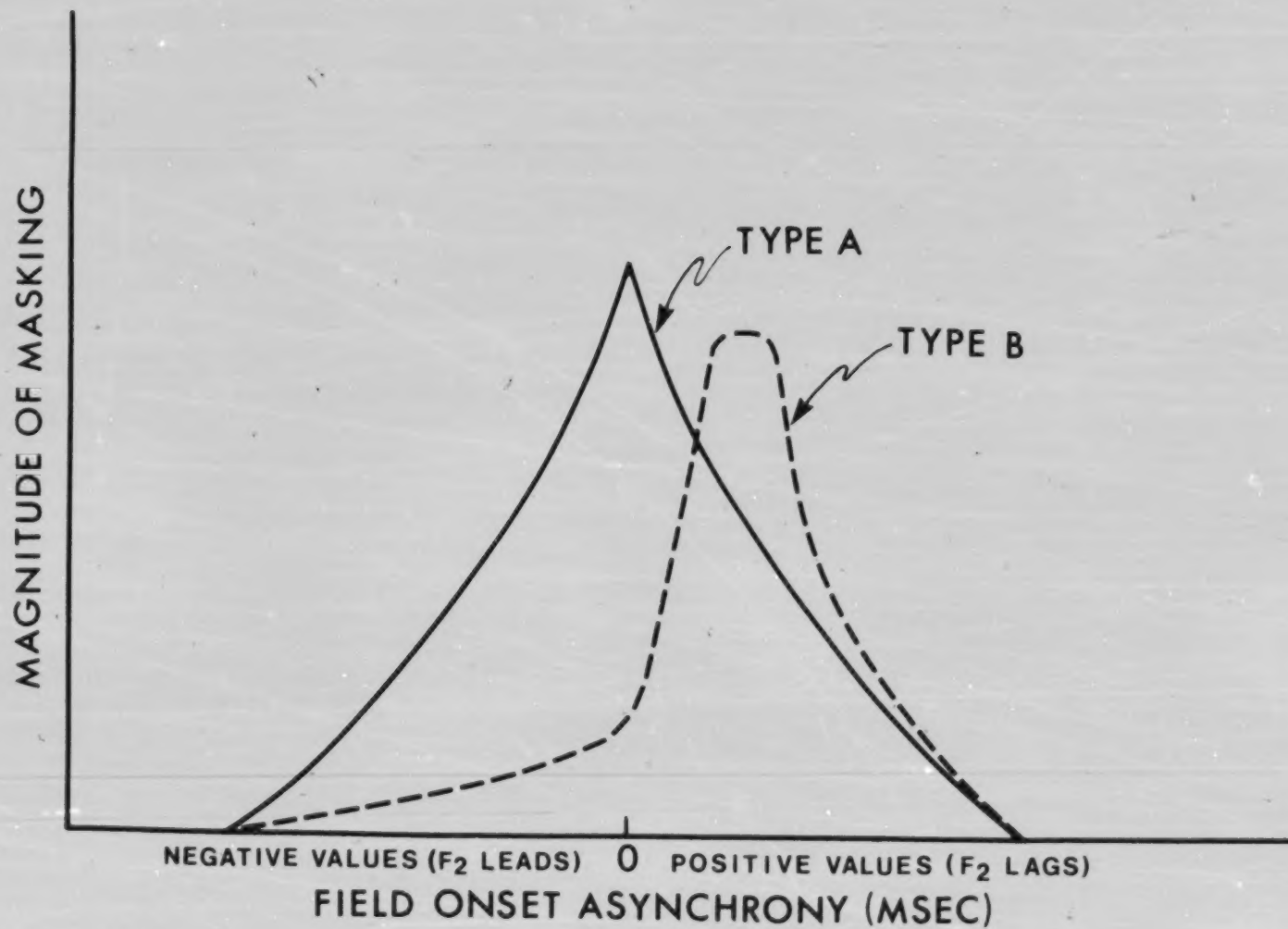
We have identified two contrasts that might afford a suitable basis for organizing the data of relevance to Two Field Theory, these are the monoptic/dichoptic contrast and the power function/additive function contrast. On first glance, it is tempting to collect together and relate *prevene* and *supervene* data under the headings of "monoptic" and "dichoptic." However, this mode of organization leads too quickly to dissonance.

Suppose I wished to claim, as some have (for example, Dick, 1974), that monoptic backward masking is greater than dichoptic backward masking because the monoptic case provides more opportunities for interaction between the two fields. From a perspective of succession-of-stages for such interactions, this claim should not hold for all instances. One might guess that, beyond some range of temporal separations, the level of processing at which the two fields interact ought to be indifferent to whether the two fields are presented to the same eye or to separate eyes. Indeed, the data bear out this information-processing intuition: while under some conditions of observation, monoptic backward masking is more severe, under others, dichoptic backward masking is its equal (for example, Turvey, 1973; Erwin and Hershenson, 1974) and may even surpass it (Turvey, 1973). Consider two further cases. One might venture the claim that, monoptically, forward masking is more pronounced than backward masking (for example, Smith and Schiller, 1966); actually, whether or not this is true depends, again, on the conditions of observation. In one experiment (Turvey, 1973, Experiment 14) with monoptic presentation, it was shown that forward masking extended over a greater range than backward masking for very brief target durations (for example, 2 msec), but over a lesser range for relatively longer target durations (for example, 20 msec). Similarly, the claim that dichoptic backward masking is more pronounced than dichoptic forward masking receives support from some research, for example, Smith and Schiller, 1966; Greenspoon and Ericksen, 1968), but is contradicted by other research (for example, Kinsbourne and Warrington, 1962b).

We see that, as an organizing principle, the monoptic/dichoptic contrast leaves much to be desired. A more accommodating approach is one that collects the data together within the context of the two two-field functions. The anomalies evident in the foregoing summary of *prevene*/*supervene* asymmetries are dispelled when one recognizes that: forward masking is more pronounced than backward masking, both monoptically and dichoptically, when the conditions favoring the power function are operating; backward masking is more pronounced than forward masking, both monoptically and dichoptically, when the conditions favoring the additive function are operating.

It remains for us to identify the two major types of masking effects--effects that are defined conventionally in terms of magnitude of masking as a function of field-onset-asynchrony. Figure 10 depicts the two types. In one, the magnitude is a monotonic function of the absolute value of the onset to onset interval; in the other, masking magnitude varies curiously with the interval: it is essentially monotonic for negative values of onset asynchrony (that is, when the target field is lagging) and nonmonotonic or U-shaped for positive values (that is, when the target field is leading). Following Kolers (1962), these two types are dubbed, respectively, type A and type B.

Figure 10: Type A and type B masking effects (magnitudes are arbitrary).



Although one can only guess at this point, it may prove valuable to Two Field Theory to recognize that the range of effective negative and positive onset-to-onset intervals incorporates the following temporal concatenates: supervene, exceed, condure, invade, and prevent. Axiomatically, the comparative durations of the two fields will determine the proportion of the effective range that is taken up by a particular concatenate. At this stage we know little about how the form of the masking effect might be influenced by this variable.

#### Masking Arises at More Than One Processing Level

An inventory of temporal relations between the two fields is complemented by an inventory of spatial and structural relations. The latter's significance to our immediate concerns is that it permits a line of inquiry leading to a conclusion of some import, namely, that there are probably different kinds of masking obeying different principles and originating at different stages in the flow of visual information.

Given two visual fields, the contours contained within the fields may or may not overlap spatially, and they may or may not relate structurally. Where the contours are spatially contiguous but not overlapping, the resultant masking is designated by the term paracontrast when the mask field,  $F_2$ , leads, and by the term metacontrast when  $F_2$  follows. Where the contours of the two fields are spatially overlapping but structurally dissimilar, for example, the target field contains a letter and the mask field is a random array of black and white regions, then the masking is referred to as masking by noise. Where the contours of the two fields are spatially overlapping and related in structure--the figural features of the one are similar to the figural features of the other--then the resultant masking is referred to as masking by pattern or masking by structure (cf. Breitmeyer and Ganz, 1976). Finally, where the mask is no more than a contourless, homogeneous light field, the resultant masking is called, not surprisingly, masking by light. Of the three kinds, other things being equal, masking by light is less pronounced than masking by noise, which in turn, is less pronounced than masking by structure (cf. Glezer, Leushina, Neuskaya and Prazdnikova, 1974).

A commonplace but eminently significant observation is that where mask energy exceeds target energy, masking by light is readily evident with monoptic presentation, but it is not evident to any substantial degree with dichoptic presentation (for example, Smith and Schiller, 1966; Kietzman, Boyle, and Lindsley, 1971). In contrast, backward masking by structure occurs for both monoptic and dichoptic viewing (for example, Smith and Schiller, 1966; Turvey, 1973). One suspects that masking by noise compromises masking by light and masking by structure. Nevertheless, we recognize that a noise mask whose grain is fine in contrast to the contours of the target, behaves very much like a homogeneous field--which is to say that it is relatively impotent as a dichoptic masker (Turvey, 1973). From these observations, we infer that with respect to a contoured target field, a contoured mask can exert a significant influence centrally, but a noncontoured mask field cannot. The influence of the latter is limited to earlier, more peripheral stages of processing.



The gist of the matter is given in Turvey's Experiment 4 (1973). A briefly exposed (say, <10 msec) field  $F_1$ , and a similarly briefly exposed and structurally-related mask field,  $F_2$ , are presented to separate eyes, with  $F_2$  lagging. A higher-energy noise field,  $F_3$ , that cannot dichoptically mask  $F_1$ , is presented shortly thereafter on the same eye as the structurally related mask field. In brief,  $F_2$  follows  $F_1$  dichoptically, and  $F_3$  follows  $F_2$  monoptically. The perceptual outcome of this complicated configuration of fields is singularly straightforward:  $F_1$  can be seen and identified against the unimpeding background of  $F_3$ ;  $F_2$  is notably absent both phenomenally and effectively.

The outcome, although straightforward, is curious and provocative. It implies that  $F_3$  prohibited or impeded the processing of  $F_2$  at a stage prior to that at which dichoptic masking originates. This is to say that  $F_2$ , which could inhibit the more central processing of  $F_1$ , never in fact "reached" the more central processing phases; in consequence, the perceptibility of  $F_1$  was unhindered. Evidently, masking phenomena are not all of one kind, nor do they arise all at one processing site.

#### Integration and Interruption as Image-Language Concepts

Information-processing accounts of visual masking have been constructed from three stock ingredients: the concepts of integration and of interruption, and the concept of iconic storage. By now it is manifestly plain that iconic storage is construed as an image in the sense of a representation that closely mimics the retinal image. Consequently, a major perspective on two-field interactions is that of two internal "images" relating to each other in ways detrimental to the perception of one or the other field, or both. Within this perspective, that we will refer to as the image-language perspective, the concepts of integration and interruption are defined as functions over internal "images." It is this perspective that defines the domain of the current section; in the section that follows, these concepts will be viewed less globally in the language of "channels" detecting features and/or spatial frequencies.

When two visual fields are presented in rapid succession, the resultant perceptual failure may be due to the lack of fine temporal resolution in the visual processing system. This is the gist of the concept of integration: it implies that two fields are treated as one--as a single package--at some subsequent processing stage. The perceptual impairment may be attributed to a confusion of features or contours or to a change in the minimum acuity requirements (Eriksen and Collins, 1965). It may also be due to summation of the luminances of the two fields. Luminance summation would reduce the contrast between the contours within the target field and their background, thereby impairing detection and identification (for example, Thompson, 1966). There is yet another way in which integration may be conceptualized. It is not so much that the two internal "images" combine, but that one field, the later one, overtakes and cancels, or reduces, the neural response to the other. At all events, whatever the manner of integration, the processing consequence is the same: at some stage processing will have to be performed on a degraded representation; we may say, on a degraded icon.

The latter remark is important because it provides a criterion for separating, conceptionally, integration from interruption. Fundamentally, interruption, as a proposed mechanism for masking, takes as its departure point clear "images," and it conceptualizes the perceptual impairment as resulting from the failure to convert fully an iconic representation into a schematic representation. We will identify two reasonably contrasting forms of the interruption concept. In one, an aftercoming field is said to erase or eliminate the iconic record of the leading field. In this view, the interruption concept shares with the integration concept the idea that masking is the direct effect of one internal representation on another.

The second form of the interruption concept is best approached through a metaphor. On entering a store, a customer is treated by the clerk as completely as possible. Unfortunately, if a second customer arrives close on the heels of the first, this shortens the amount of time that the clerk can devote to the first. Where customers enter the store on an aperiodic schedule, the treatment that any one customer receives is a function of whether, and how close, a second customer follows. In this metaphor, the clerk is a central processor and the customers are visual fields. Through the metaphor, we understand that if one field follows another after some delay, processing is assumed to have occurred during the delay, but it is terminated or interfered with by the following field. The clerk-customer metaphor (Kolars, 1968) emphasizes not the effect of one field on another, but the effect of one field, the lagging field, on the central processor's activities. Furthermore, and significantly, the metaphor suggests that the record of the first field may persist even though it is no longer being processed; after all, a first customer is not eliminated by the arrival of a second. In these two accounts, the second version of the interruption concept differs from the first. Nevertheless, for either version, it is apparent that interruption is more befitting to the prevene or backward masking condition, than the supervene condition--unlike integration which is befitting to both.

There has been a tendency to consider the two masking mechanisms of integration and interruption as mutually exclusive, and experiments seeking to demonstrate that masking was due to one rather than the other have been reported (for example, Coltheart and Arthur, 1972). It is more prudent and judicious to bet on an account of masking that is a hybrid of the two (for example, Spencer and Shuntich, 1970; Scheerer, 1973). We recall that in the second section (Isolating the Brief Icon), Averbach and Coriell (1961) saw fit to speak of summation and erasure as yoked causes of their masking functions. One most obvious benefit of a hybrid theory is that it gives a rationale for why backward masking by structure tends to be more severe and more temporally extensive than forward masking by structure (see Scharf and Lefton, 1970; Spencer and Shuntich, 1970); forward masking has but one ingredient (integration); backward masking has two (integration and interruption).

### Integration and Interruption as Channel-Language Concepts

What defines a period of integration? When our thinking is couched in the language of images, we are likely to suppose that it is defined either by some overall field property, such as total energy, or by some internal process that, figuratively speaking, partitions time into discrete frames, or moments of a single size.

Consider the well-established finding of a reciprocity law that relates intensity and duration of visual stimulation. According to this law, a given perceptual effect depends only on total energy and is indifferent to the relative contribution of duration and intensity to that total, up to some critical duration. This critical duration may be said to define the integration period. Curiously, the period of integration is not invariant: it is less for brightness discrimination than for identification and acuity judgments (Kahneman, 1967); it is less when the response measure is latency of detection, than when it is the frequency of seeing a target or signal detectability (Bruder and Kietzman, 1973). The critical duration for brightness discrimination can differ from that for form identification by as much as 150 msec. Observations such as these are discouraging to the theorist who aspires to identify a single period of integration that dominates the processing of visual information. Quite to the contrary, it appears that different periods of integration are associated with different properties of stimulation.

If, indeed, there are many moments rather than one, how might they relate? One answer that comes quickly to mind is that they are nested. Given a particular property, there is an interval over which the perceptual system will integrate stimulation relevant to that property; the upper limit on this interval is contained within the interval over which integration occurs for another, different property. Likewise, the upper limit on this latter moment is contained within the range of the next, and so on. In a phrase, the idea of nested moments conveys the principle of different processes operating simultaneously at different rates. This principle is at the heart of recent channel-language accounts of masking (Turvey, 1973; Breitmeyer and Ganz, 1976).

To apply this principle, we have only to imagine the mapping from retina to cortex as a set of channels organized in parallel and relatively independent in function. The sensitivity of these channels could be described in terms of features, in which case we would remark that the lower the order of the feature on some scale of informational complexity, the shorter is its integration period or equivalently, the more rapidly is it detected (cf. Turvey, 1973).

The principle is illustrated all the more clearly in spatial-frequency terms. For here we can speak unequivocally of a continuum from (very) low to (very) high spatial frequencies and propose, without too much difficulty, the existence of spatial-frequency channels (see Sekuler, 1974). There is a growing suspicion that the higher the spatial frequency, the slower the channel operating time (Breitmeyer and Ganz, 1976). Or, in the language of moments, the higher the spatial frequency, the longer the period of

integration. There is a further advantage to the spatial frequency terminology; and it is that different spatial frequencies contribute differently to the content of perception. To be less than perfectly precise, low spatial frequency channels are sensitive to change and contribute to localization; intermediate to high spatial frequency channels are sensitive to relational contour information and contribute to form identification; very high spatial frequency channels permit the accentuating of contours and hence contribute to the clarity of the percept (see Breitmeyer and Ganz, 1976).

An even less sensitive but exceptionally useful distinction can be drawn: given a brief exposure, transient channels respond to the onset of the exposure, and sustained channels respond during and beyond the exposure. The latter are said to support the identification of form, and their responsiveness to stimulation is both slower and more persistent than that of transient channels. This transient/sustained distinction (cf. Breitmeyer and Ganz, 1976) is reminiscent of that drawn between two visual systems (cf. Trevarthan, 1968), one responsible for determining where a certain thing is in the field of view, and the other responsible for determining what it is. In the relation between the two classes of channels, we observe that transient channel activity may inhibit sustained channel activity (Singer and Badworth, 1973). Given two brief exposures in rapid succession, the short-latency transient channel activity induced by the second will impose on the longer-lasting sustained activity induced by the first. Where the output of sustained channels is significant to identification, the inhibition of their activity by transient channels will result in an incomplete identification. In short, the transient channel/sustained channel interaction suggests an interruption mechanism for backward masking (cf. Breitmeyer and Ganz, 1976).

It must be remarked, therefore, and in summary, that in the case of the concepts of integration and interruption there is a "channel language" that describes them. At least in part, integration can be thought of as time-sharing within a channel, and interruption as the inhibition of the activity in one class of channels by the activity in another.

### Three Interpretations of Nonmonotonicity

Let us exercise the elementary principles and intuitions of the immediately preceding sections through the explication of three, reasonably separate interpretations of Type B structure masking. The first is expressed in image language, the second is expressed in a mix of the two languages, and the third is stated purely in terms of channels. The problem for each is that Type B effects are evident at low mask to target energy ratios, while Type A effects are evident at high mask to target energy ratios (see Figure 11); the Type B effect, we note in passing, has been the traditional bete noir of Two Field Theory (Kahneman, 1968).

The first--an elegantly simple interpretation--is due to Spencer and Shuntich (1970). Their central thesis is that in the prevenue condition one needs to consider the degree of  $F_1$  processing prior to the advent of  $F_2$ , and the degree of  $F_1$  processing subsequent to  $F_2$ . The account assumes that the iconic representation of  $F_1$  is immediately established. When the masker,  $F_2$ ,



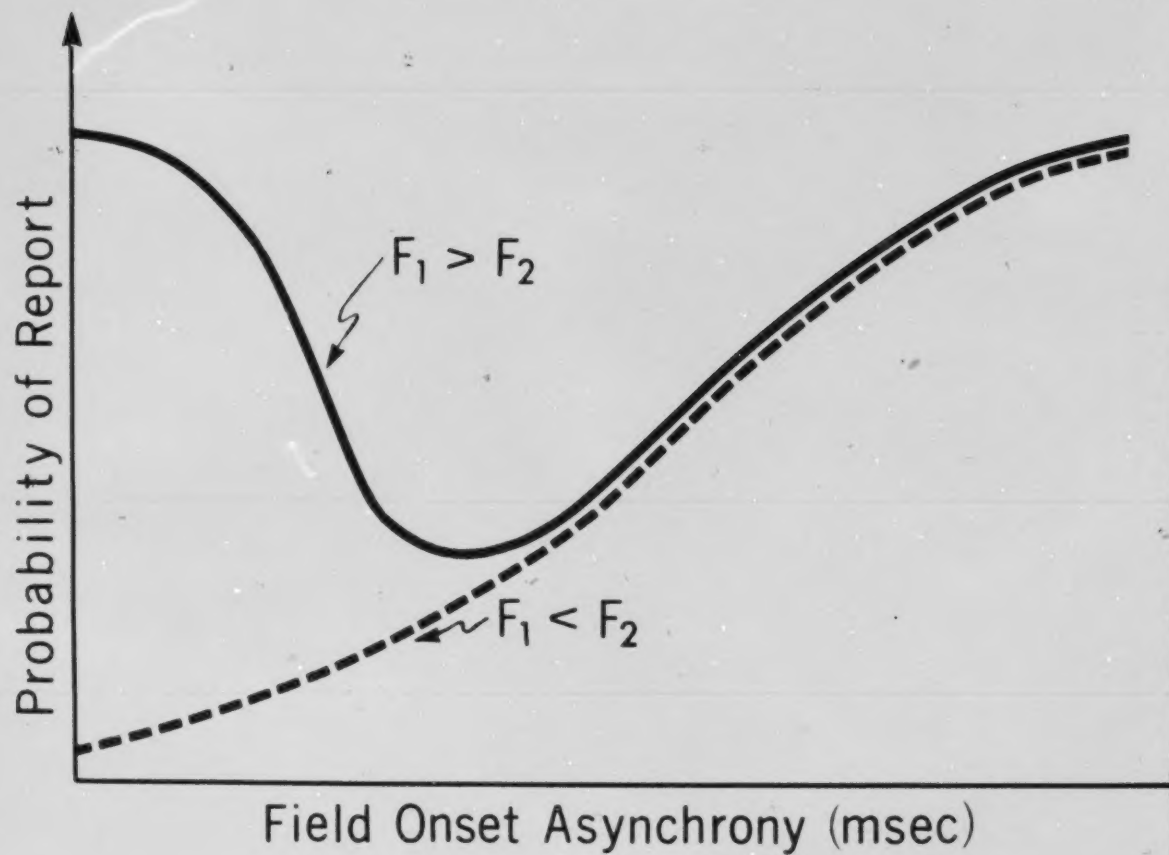


Figure 11: The type of masking is affected by the energy relation between the two successive fields.

is of considerably higher energy, the claim is that it eliminates the  $F_1$  icon from further analysis. As a consequence, the perceptibility or reportability of  $F_1$ , will be a function simply of pre- $F_2$  processing and should increase monotonically with the onset asynchrony between the two fields. In contrast, when the energy differential favors the target field,  $F_1$ , the icon of  $F_1$  may persist (albeit less than perfectly legible since  $F_1$  and  $F_2$  will have integrated) beyond the advent of  $F_2$ . In this case, therefore, reportability of the target is a function of both pre- and post- $F_2$  processing. Inasmuch as the icon decays with time, the capability of the later, lower-energy  $F_2$  to perturb the iconic record of  $F_1$ , is enhanced at greater temporal separations of the two fields. It follows that with increasing delay of  $F_2$ , an increasing order of pre- $F_2$  processing is accompanied by a decreasing order of post- $F_2$  processing. In this shifting balance between pre- and post- $F_2$  processing lies the genesis of nonmonotonicity (Spencer and Shuntich, 1970).

The starting point for the second interpretation (Turvey, 1973) is that the light to an eye is analyzed initially by a set of operationally parallel and independent peripheral channels. Each channel is said to be sensitive to a different kind of arrangement in the light and each is said to respond to its preferred arrangement at its own particular rate. The outputs of these channels are registered in a set of central "mini-stores," and it follows that this registration occurs asynchronously, that is, some outputs are registered ahead of others. When presented in close succession, two fields will "occupy" the same peripheral channels to the extent that they are structurally similar. Double occupancy of a peripheral channel favors the greater energy field; in other words, what is entered into the ministore allied to a channel is more likely to be a datum of the field with greater energy. Thus, when the onsets of the two fields are virtually simultaneous and the mask field,  $F_2$ , has the greater energy, the information in the central ministores will be predominantly mask information and, in consequence, target identification will be minimal. If  $F_1$  has the greater energy, then the situation is reversed.

The peripheral channels can be said to detect context independent properties that are used to reconstitute the wholistic character of the patterned light at the eye. The claim is for a process of construction or synthesis that takes as its ingredients the outputs of the peripheral channels, that is, the data in the ministores. The relation between the activity in the peripheral channels and the central synthesis is described as concurrent and contingent, which is meant to imply that the synthesis proceeds in parallel with the detection of context independent properties, and is predicated upon them. What is synthesized is the icon; from this representation is abstracted context dependent properties necessary to identification.

There remain to be identified two variants of interruption that this account exploits before we can speak to nonmonotonicity. In the first place, there is the idea that the entry in a ministore allied to a channel is changed by an entry from the same channel. Given two structurally related fields separated by an interval that exceeds the integration period of a particular channel, the record from the first field's occupancy of the channel will be replaced by the record of the second field's occupancy of the

channel. We may dub this the "replacement principle" and observe that it is an instance of a general rule, namely, that when there is competition for central mechanisms, the victor is likely to be the field that is presented second. Centrally, order is more significant than energy, and a lagging exposure will dominate one that leads.

In like fashion, the other variant of interruption is a manifestation of the replacement principle; its location, however, is further upstream in the central flow of information. Given two structurally related fields that are separated by an interval in excess of the synthesizing period, the processes performed on the iconic representation of  $F_1$  will be interrupted by the advent of the iconic representation of  $F_2$ . This latter form of interruption is at the conceptual level of the clerk in the clerk-customer metaphor;  $F_1$  is replaced, as the focus of the clerk's attention, by  $F_2$ .

In terms of the above, let us see what should transpire when  $F_1$  energy exceeds that of  $F_2$ . At simultaneity, as already remarked, the two fields will occupy the set of channels they share in common, and  $F_1$  will occlude  $F_2$ . As the onset asynchrony is increased, the peripheral processing of the two fields will be only partially overlapping temporally. Where the two fields time-share a particular channel,  $F_1$  will dominate; where they do not, that is, where a channel has finished processing  $F_1$  before it is occupied by  $F_2$ ,  $F_2$  will dominate, given the replacement principle. Consequently, at very brief separations, information on both  $F_1$  and  $F_2$  will be included in the synthesis, a composite icon will result, and the fidelity of  $F_1$  will be marred. In very few words we can describe the balance of the story: as the interval between  $F_1$  and  $F_2$  increases, the energy superiority of  $F_1$  becomes less important, and the fact that  $F_2$  is second in arrival becomes more important. In this manner a lower energy  $F_2$  that fails to mask at very brief delays, masks significantly at longer delays. Of course, with ever increasing delays, the masking capability of the lagging  $F_2$  decreases, owing to the increasing likelihood that  $F_1$  has been coded into a nonmaskable, short-term store. To conclude, if  $F_1$  was less in energy than  $F_2$ , then the Type A effect would arise, because  $F_2$  would impede perception at very brief delays, due to its energy superiority, and at relatively longer delays, due to its status as the lagging member of the pair.

We come, finally, to the third interpretation. It is due to Breitmeyer and Ganz (1976) and it shares the following ideas with the interpretation just described: first, the idea of parallel peripheral channels detecting different properties of stimulation--precisely, spatial frequencies, at different rates; second, the idea that integration may occur within peripheral channels and within the process of central synthesis. These two forms of channel-language integration deserve special labels. The following are suggested: integration through channel time-sharing and integration through common synthesis, respectively. Of course, both modes of integration may be conceptualized as the adding of noise. We may remark that, for either mode, the integration period, in the view of Breitmeyer and Ganz (1976), is a function of the constituent spatial frequencies common to the two fields; for progressively higher spatial-frequency components, the integration period is progressively longer.

According to this third interpretation, monotonic Type A masking effects are the consequence of either or both modes of integration, but Type B effects are due to a third mechanism, the inhibition of sustained channels by transient channels. The argument is that inhibition through transient channel activity is optimal when the activity in transient and sustained channels is contemporaneous. Because of the latency difference alluded to previously, a contemporary state is achieved when, in the prevence condition,  $F_2$  is delayed by several tens of milliseconds or more, relative to the onset of  $F_1$ . Yet, the delay for optimal inhibition ought to be a function of the observer's task. According to a previous claim, identifying a form contained in  $F_1$  requires intermediate to high spatial frequencies, and the injection of this information into the central synthesis occurs sooner than the higher spatial frequencies necessary to the making of acuity judgments. Given this understanding, in order to maximize masking, the delay in the onset of  $F_2$  relative to  $F_1$  would be predictably less for form identification, than for contour resolution.

Let us not forget that the original problem was to explain why there is a transition from Type A to Type B effects accompanying a transition from high mask-to-target energy ratios to low mask-to-target energy ratios. In the style of the previous two interpretative schemes, it is necessary for the current scheme to assume that the greater energy field is perceptually dominant at brief separations. In this particular case, the field of greater energy dominates the activity of sustained channels. In addition, the current interpretation claims that when the mask field increases in energy relative to the target field, the sustained channels activity increases considerably, while that of the transient channels increases only slightly. There is, in consequence, a proposed shift in the relative activity of the two kinds of channels with a transition in the  $F_2$ -to- $F_1$  energy ratio; and it is this shift that is offered as the solution to the aforementioned problem (Breitmeyer and Ganz, 1976).

#### N-Field Theory?

We conclude the second part of this chapter with the explanation of Type A and Type B masking effects. The survey of two-field phenomena and of the related theory has been selective but, it is hoped, sufficient. Ideally, the major issues have been touched. Comprehending the perceptual consequences of two independent fields presented in rapid succession to the same and fixed point of observation is a task that is far from complete. Furthermore, the stock of concepts apropos the facts of One Field Theory need enriching in order to accommodate the facts that face a Two Field Theory.

There is, of course, no a priori reason why we should restrict our attention to the phenomena related to one- or two-field exposures. Three-field phenomena are to be found in the literature and we described one such phenomenon above (masking arises at more than one processing level). These phenomena, not surprisingly, are even more perplexing than their one- and two-field counterparts. Fields may appear and disappear at random. A field that is perceived poorly in a two-field situation may be perceived significantly better in a three-field situation (for example, Sperling,



1970). A field that is impervious to a leading or a lagging mask in the independent two-field cases, may succumb to these individually impotent mask fields in the three-field case (Uttal, 1969)--an effect that, we have good reason to believe, is not to be understood in terms of a simple summation of latent masking influences (Turvey, 1973). In short, the perceptual consequences of three-field interactions look inordinately more intricate than those of the two-field case. And one can only wonder at the form and complexity of an N-Field Theory. What are the perceptual consequences of  $n$  successive, separate fields? This question is of central concern to the contemporary theory of visual processing and a little history reveals why this is so.

Two notions of considerable antiquity bear significantly on our conceptions of iconic and schematic memory: the concept of the retinal image, and the camera obscura as a metaphor for the eye. Conventionally, the retinal image is described as a far from adequate simulation of the environmental arrangement at which the eye is directed (see Section A, Visual Information Processing in Historical Perspective). The eye-as-a-camera theme permits further elaboration: insofar as the retina is likened to a photographic plate, the light falling upon it must be limited to but a brief moment--just sufficient for the forming of an image. Any longer and the image will blur, a consequence that also would be incurred by moving the eye or the environment. In brief, our point of view is that the eye is a device that captures static bidimensional images, and the visual system is a mechanism that analyzes them (see Turvey, 1977).

From this historical perspective, it is no great leap to the conception of iconic and schematic memory as static representations, for they are, after all, correlates of the retinal image. This conception might be reinforced by indications that the icon, for example, is a record of the anatomical pattern of retinal cells that are excited. The source of iconic persistence appears to be localized primarily in the photoreceptors (Sakitt, 1975), and masking, which is pronounced when the target and mask share retinal coordinates (that is, when they stimulate the same receptor elements), is virtually absent when the target and mask coincide only in environmental coordinates (Davidson, Fox and Dick, 1973).

To illustrate the influence of the legacy even further, consider the situation in which an observer experiences an event, that is, a change wrought over an object; most obviously, the optical information relevant to such an experience is realized over time. From the traditional perspective, the experience or "perception" of an event is a deduction from a sequence of static arrangements--in contemporary parlance we would say that it is a deduction from a sequence of iconic and/or schematic memories. The preceding sentence characterizes the most eminent feature of the visual processing-research and theory that has been the focus of this paper: the analysis of visual processing into discrete temporal cross-sections perpendicular to the flow of optical information (Turvey, 1977). The significance of the above question concerning the perceptual consequences of  $n$  successive separate fields can now be appreciated.

Yet we must comment, in these concluding remarks, on the possibility that the above question is ill conceived. The analysis of visual processing as discrete temporal cross-sections perpendicular to the optical flow is well motivated from historical considerations. Nevertheless, it is a most unattractive orientation when one looks quizzically and seriously at how a continuous optical flow might be decomposed, and how discrete samples might be analyzed or integrated to reveal the style of change (Turvey, 1977). Given a moving observer and an environment that is undergoing change, for example, what is the mechanism which procures, in stroboscopiclike fashion, discrete static samples of the optical flow field, given a moving observer and an environment that is undergoing change? Aside from the limiting case of a stationary observer moving his eyes over a frozen environment--a case in which successive fixations might be hypothesized as the source of discontinuous, static samples--there is no sensible mechanism to which we can turn.

If discrete samples are not realizable, then iconic or schematic memories, as correlates of retinal "snapshots," cannot be the informational support for the perception of events. However, if the informational support for event perception is the continuously transforming optical flow, how might event perception be characterized? Following Gibson's (1966) lead, we may regard the perception of events as dependent upon the detection of invariants over time, rather than on the perception of static forms. With respect to events, we may define a transformational invariant as that information specific to the style of change that is preserved over different structures "supporting" the change, and a structural invariant as that information specific to the object structure that is preserved over the styles of change in which the object participates (see Pittenger and Shaw, 1975; Shaw and Pittenger, 1976; Turvey, in press). To speak about the perception of an event as depending on the detection of invariants that are specific to the structure of the event, is to introduce an odd perspective on the theory of visual processing. The oddity arises from the denial of an equivocal, imperfect relation between the structure of the event and the light as structured by the event--that is, from a denial of the fundamental precept of indirect realism (see Part A, Visual Processing as a Version of Indirect Realism). If the optical support for visual processing is informationally rich and specific to the properties of the environment, then visual processing need not be a matter of inference or of elaboration from memory, contrary to Helmholtz's understanding (see Part A, Visual Information Processing in Historical Perspective) and contrary to current visual processing schemes (see Part A, Visual Information Processing: A Preliminary Portrayal).

From these concluding remarks one can intuit that a theory of visual processing in the perspective of direct realism (see Shaw and Bransford, 1976; Turvey, in press) will differ markedly from the familiar theory with which we have been most concerned in the present chapter--namely, the theory of visual processing in the perspective of indirect realism.

# REFERENCES

- Avant, L. L. and P. M. Lyman. (1975) Stimulus familiarity modifies perceived duration in prerecognition visual processing. J. Exper. Psychol.: Human Percept. Perform. 1, 205-213.
- Avant, L. L., P. J. Lyman, and J. Antes. (1975) Effects of stimulus familiarity upon judged visual duration. Percept. Psychophys. 17, 253-262.
- Averbach, E. and A. S. Coriell. (1961) Short-term memory in vision. Bell System Technical Journal 40, 309-328.
- Blakemore, C. and F. W. Campbell. (1969) On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. J. Physiol. 203, (London) 237-260.
- Bennett, I. F. (1971) Spatial effects of visual selective attention. Human Performance Center Technical Report No. 23, The University of Michigan, Ann Arbor.
- Besner, D., J. K. Keating, L. J. Cake, and R. Maddigan. (1974) Repetition effects in iconic and verbal short-term memory. J. Exper. Psychol. 102, 901-902.
- Breitmeyer, B. G. and L. Ganz. (1976) Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression and information processing. Psychol. Rev. 83, 1-36.
- Briggs, G. C. and M. Kinsbourne. (1972) Visual persistence as measured by reaction time. Quart. J. Exper. Psychol. 24, 318-325.
- Broadbent, D. E. (1958). Perception and Communication. (New York: Pergamon Press).
- Broadbent, D. E. (1971) Decision and Stress. (London: Academic Press).
- Bruder, G. E. and M. L. Kietzman. (1973) Visual temporal integration for threshold, signal detectability and reaction time measures. Percept. Psychophys. 13, 293-300.
- Cermak, G. W. (1971) Short-term recognition memory for complex free-form figures. Psychon. Sci. 25, 209-211.
- Chow, S. L. and B. B. Murdock. (1975) The effect of a subsidiary task on iconic memory. Mem. Cog. 3, 678-688.
- Chow, S. L. and B. B. Murdock. (1976) Concurrent memory load and the rate of readout from iconic memory. J. Exper. Psychol.: Human Percept. Perform. 2, 179-190.
- Clark, S. E. (1969) Retrieval of color information from the preperceptual storage system. J. Exper. Psychol. 82, 263-266.
- Clowes, M. (1971) On seeing things. Artificial Intelligence 2, 79-112.
- Cohen, R. L. and B. S. Johansson. (1967) The activity trace in immediate memory: A reevaluation. J. Verbal Learn. Verbal Behav. 6, 139-143.
- Coltheart, M. and B. Arthur. (1972) Evidence for an integration theory of visual masking. Quart. J. Exp. Psychol. 24, 262-269.
- Craik, F. I. M. and E. Tulving. (1975) Depth of processing and the retention of words in episodic memory. J. Exp. Psychol.: General 104, 267-294.
- Davidson, M. L., M. J. Fox and A. O. Dick. (1973) Effects of eye movements on backward masking and perceived location. Percept. Psychophys. 14, 110-116.

- Dick, A. O. (1969) Relations between the sensory register and short-term storage in tachistoscopic recognition. J. Exp. Psychol. 82, 279-284.
- Dick, A. O. (1970) Visual processing and the use of redundant information in tachistoscopic recognition. Canad. J. Psychol. 24, 113-141.
- Dick, A. O. (1971) On the problem of selection in short-term visual (iconic) memory. Canad. J. Psychol. 25, 250-263.
- Dick, A. O. (1972) Visual hierarchical feature processing: The relation of size, spatial position and identity. Neuropsychologica 10, 171-177.
- Dick, A. O. (1974) Iconic memory and its relation to perceptual processing and other memory mechanisms. Percept. Psychophys. 16, 575-596.
- Doost, R. and M. T. Turvey. (1971) Iconic memory and central processing capacity. Percept. Psychophys. 9, 269-274.
- Efron, R. (1973) An invariant characteristic of perceptual systems in the time domain. In Attention and Performance IV, ed. by S. Kornblum. (New York: Academic Press).
- Eriksen, C. N. and R. L. Colgate. (1971) Selective attention and serial processing in briefly presented visual displays. Percept. Psychophys. 10, 321-326.
- Eriksen, C. W. and J. F. Collins. (1965) A reinterpretation of one form of backward and forward masking in visual perception. J. Exp. Psychol. 70, 343-351.
- Eriksen, C. W. and J. F. Collins. (1969) Temporal course of selective attention. J. Exp. Psychol. 80, 254-261.
- Eriksen, C. W. and J. F. Collins. (1967) Some temporal characteristics of visual pattern perception. J. Exp. Psychol. 74, 476-484.
- Eriksen, C. W. and J. E. Hoffman. (1972) Temporal and spatial characteristics of selective encoding from visual displays. Percept. Psychophys. 12, 201-204.
- Erwin, D. E. (1976) Further evidence for two components in visual persistence. J. Exper. Psychol.: Human Percept. Perform. 2, 191-209.
- Erwin, D. E. and M. Hershenov. (1974) Functional characteristics of visual persistence predicted by a two factor theory of backward masking. J. Exp. Psychol. 103, 249-254.
- Estes, W. K. and H. A. Taylor. (1966) Visual detection in relation to display size and redundancy of critical elements. Percept. Psychophys. 1, 9-16.
- Fryklund, I. (1975) Effects of cued-set spatial arrangement and target background similarity in the partial-report paradigm. Percept. Psychophys. 17, 375-386.
- Gibson, J. J. (1966) The Senses Considered as Perceptual Systems. (Boston: Houghton Mifflin).
- Glezer, V. D., L. I. Leushina, A. A. Nevskaya, and N. V. Prazdnikova. (1974) Studies on visual pattern recognition in man and animals. Vision Res. 14, 555-583.
- Glucksberg, S. and S. Balagura. (1965) Effects of repetition and intra-array similarity upon very short-term visual memory. Paper presented at meetings of Psychonomic Society, Chicago.
- Golañi, I. (in press) Homeostatic motor processes in mammalian interactions: A choreography of display. In Perspectives in Ethology, ed. by P. Bateson and P. Lopfež. (London: Plenum Press).
- Greenspoon, T. S. and C. S. Eriksen. (1968) Interocular non-independence.



- Percept. Psychophys. 3, 93-96.
- Gregory, R. L. (1970) The Intelligent Eye. (New York: McGraw-Hill).
- Guzman, A. (1969) Decomposition of a visual scene into three-dimensional bodies. In Automatic Interpretation and Classification of Images, ed. by A. Graselli. (New York: Academic Press).
- Haber, R. N. (1969) Information processing analyses of visual perception: An introduction. In Information Processing Approaches to Visual Perception, ed. by R. N. Haber. (New York: Holt, Rinehart & Winston).
- Haber, R. N. (1971) Where are the visions in visual perception. In Imagery, ed. by S. Segal. (New York: Academic Press).
- Haber, R. N. and L. Standing. (1969) Direct measures of short-term visual storage. Quart. J. Exp. Psychol. 21, 43-54.
- Haber, R. N. and L. Standing. (1970) Direct estimates of apparent duration of a flash followed by visual noise. Canad. J. Psychol. 24, 216-229.
- Halle, M. and K. Stevens. (1962) Speech recognition: A model and a program for research. IRE Transactions on Information Theory, IT-8, 155-159.
- Hebb, D. O. (1949) The Organization of Behavior. (New York: Wiley).
- Hebb, D. O. (1961) Distinctive features of learning in the higher animal. In Brain Mechanisms and Learning, ed. by J. F. Delagresnaye. (New York: Oxford University Press).
- Helmholtz, H. von. (1925) Treatise on psychological optics, ed. and trans. from the 3rd German ed. (1909-1911) by J. P. Southall. (Rochester, N.Y.: Optical Society of America).
- Henderson, L. (1972) Spatial and verbal codes and the capacity of STM. Quart. J. Exp. Psychol. 24, 485-495.
- Hochberg, J. (1974) Higher-order stimuli and inter-response coupling in the perception of the visual world. In Perception: Essays in Honor of J. J. Gibson, ed. by R. B. MacLeod and H. L. Pick. (Ithaca: Cornell University Press).
- Holding, D. (1970) Guessing behavior and the Sperling store. Quart. J. Exp. Psychol. 22, 248-256.
- Hubel, D. H. and T. N. Wiesel. (1967) Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. J. Neurophysiol. 30, 1561-1573.
- Hubel, D. H. and T. N. Wiesel. (1968) Receptive fields and functional architecture of monkey striate cortex. J. Physiol. 195, 215-243.
- Kahneman, D. (1967) Temporal effects in the perception of light and form. In Models for the Perception of Speech and Visual Form, ed. by W. Wathen-Dunn. (Cambridge: MIT Press).
- Kahneman, D. (1968) Method, findings and theory in studies of visual masking. Psychol. Bull. 70, 404-426.
- Kinsbourne, M. and E. K. Warrington. (1962 a) The effect of an aftercoming random pattern on the perception of brief visual stimuli. Quart. J. Exp. Psychol. 14, 223-224.
- Kinsbourne, M. and E. K. Warrington. (1962 b) Further studies on the masking of brief visual stimuli by a random pattern. Quart. J. Exp. Psychol. 14, 235-245.
- Kietzman, M. L., R. C. Boyle, and D. B. Lindsley. (1971) Perceptual masking: Peripheral vs central factors. Percept. Psychophys. 9, 350-351.
- Kolers, P. A. (1962) Intensity and contour effects in visual masking.

Vision Res. 2, 277-294.

- Kolers, P. A. (1968) Some psychological aspects of pattern recognition. In Recognizing Patterns, ed. by P. A. Kolers and M. Eden. (Cambridge: MIT Press).
- Kroll, N. E. A., T. Parks, S. R. Parkinson, S. L. Bieber, and A. L. Johnson. (1970) Short-term memory while shadowing: Recall of visually and of aurally presented letters. J. Exp. Psychol. 85, 220-224.
- Lowe, D. G. (1975) Processing of information about location in brief visual displays. Percept. Psychophys. 18, 309-316.
- Maturana, H. R., J. Y. Lettvin, W. S. McCullough, and W. H. Pitts. (1960) Anatomy and physiology of vision in the frog. J. Gen. Physiol. 43(S), 129-171.
- Melton, A. W. (1963) Implications of short-term memory for a general theory of memory. J. Verbal Learn. Verbal Behav. 2, 1-21.
- Merluzzi, T. V. and N. F. Johnson. (1974) The effect of repetition on iconic memory. Quart. J. Exp. Psychol. 26, 266-273.
- Mewhort, D. J. K. (1972) Scanning, chunking and the familiarity effect in tachistoscopic recognition. J. Exp. Psychol. 93, 69-71.
- Meyer, G. E., R. Lawson, and W. Cohen. (1975) The effects of orientation-specific adaptation on the duration of short-term visual storage. Vision Res. 15, 569-572.
- Mitchell, D. C. (1972) Short-term visual memory and pattern masking. Quart. J. Exp. Psychol. 24, 294-405.
- Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).
- Norman, D. A. (1968) Toward a theory of memory and attention. Psychol. Rev. 75, 722-736.
- Norman, D. A. and D. G. Bobrow. (1975) On data-limited and resource-limited processes. Cog. Psychol. 7, 44-64.
- Novik, N. (1974) Development studies of backward visual masking. Unpublished Ph.D. thesis, University of Connecticut.
- Ornstein, R. E. (1969) On the Experience of Time. (Middlesex, England: Penguin Books).
- Phillips, W. A. (1974) On the distinction between sensory storage and short-term visual memory. Percept. Psychophys. 16, 282-290.
- Phillips, W. A. and A. D. Baddeley. (1971) Reaction time and short-term visual memory. Psychon. Sci. 22, 73-74.
- Pittenger, J. B. and R. E. Shaw. (1975) Aging faces as viscal-elastic events: Implications for a theory of non-rigid shape perception. J. Exp. Psychol:Human Percept. Perform. 1, 374-382.
- Pollack, I. (1973) Interaction effects in successive visual displays: An extension of the Eriksen-Collins paradigm. Percept. Psychophys. 13, 367-373.
- Posner, M. I. (1969) Abstraction and the process of recognition. In Psychology of Learning and Motivation vol. 3, ed. by G. H. Bower and J. T. Spence. (New York: Academic Press).
- Robinson, D. N. (1968) Some properties of visual short-term memory. Percept. Mot. Skills 27, 1155-1158.
- Ross, J. and J. H. Hogben. (1974) Short-term memory in stereopsis. Vision Res. 14, 1195-1201.
- Rudov, M. (1966) Dimensionality of human information storage.

- J. Exp. Psychol. 71, 273-281.
- Sakitt, B. (1975) Locus of short-term visual storage. Science 190, 1318-1319.
- Sanders, A. F. (1968) Short-term memory for spatial positions. Psychologic 23, 1-15.
- Scarborough, D. L. (1972) Memory for brief visual displays of symbols. Cog. Psychol. 3, 408-429.
- Scharf, B. and L. A. Lefton. (1970) Backward and forward masking as a function of stimulus and task parameters. J. Exp. Psychol. 84, 331-338.
- Scheerer, E. (1973) Integration, interruption and processing rate in backward masking. I. Review. Psychologische Forschung 36, 71-93.
- Schiller, P. H. (1965) Monoptic and dichoptic visual masking by patterns and flashes. J. Exp. Psychol. 69, 193-199.
- Sekuler, R. (1974) Spatial vision. In Ann. Rev. Psychol. (Palo Alto: Annual Reviews, Inc.).
- Shaw, R. E. and J. Bransford. (1976) Introduction: Psychological approaches to the problem of knowledge. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. E. Shaw and J. Bransford. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Shaw, R. and J. Pittenger. (1976) Perceiving the face of change in changing faces: Implications for a theory of object perception. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. E. Shaw and J. Bransford. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Singer, W. and N. Bedworth. (1973) Inhibitory interaction between X and Y units in the cat lateral geniculate nucleus. Brain Res. 49, 291-307.
- Smith, M. C. and P. H. Schiller. (1966) Forward and backward masking: A comparison. Canad. J. Psychol. 20, 339-342.
- Spencer, T. J. and R. Shuntich. (1970) Evidence for an interruption theory of backward masking. J. Exp. Psychol. 85, 198-203.
- Sperling, G. (1960) The information available in brief visual presentations. Psychol. Monogr. 74 (Whole No. 498).
- Sperling, G. (1963) A model for visual memory tasks. Human Factors 5, 19-31.
- Sperling, G. (1967) Successive approximations to a model for short-term memory. Acta Psychol. 27, 285-292.
- Sperling, G. (1970) Short-term memory, long-term memory and scanning in the processing of visual information. In Early Experience and Visual Information Processing in Perceptual and Reading Disorders, ed. by F. A. Young and D. B. Lindsley. (Washington: National Academy of Science).
- Sperling, G. (1971) Information processing from two rapidly consecutive stimuli: A new analysis. Percept. Psychophys. 9, 89-91.
- Standing, L. and F. DaPolito. (1968) Limitations of the repetition effect revealed by partial report. Psychon. Sci. 13, 297-299.
- Standing, L. G. and P. C. Dodwell. (1972) Retroactive contour enhancement: A new visual storage effect. Quart. J. Exp. Psychol. 24, 21-29.
- Stevens, S. S. (1970) Neural events and the psychological law. Science 170, 1043-1050.
- Sutherland, N. S. (1973) Intelligent picture processing. Paper presented at Conference on the Evolution of the Nervous System and Behavior,

- Florida State University, Tallahassee.
- Thomas, E. A. C. and W. B. Weaver. (1975) Cognitive processing and time perception. Percept. Psychophys. 17, 363-367.
- Thompson, J. H. (1966) What happens to the stimulus in backward masking? J. Exp. Psychol. 71, 580-586.
- Treisman, A., R. Russell, and J. Green. (1974) Brief visual storage of shape and movement. In Attention and Performance V. (New York: Academic Press).
- Townsend, V. M. (1974) Loss of spatial and identity information following a tachistoscopic exposure. J. Exp. Psychol. 98, 113-118.
- Trevarthan, C. B. (1968) Two mechanisms of vision in primates. Psychologische Forschung 31, 299-337.
- Tulving, E. and T. Y. Arbuckle. (1963) Sources of intratrital interference in immediate recall of paired associates. J. Verbal Learn. Verbal Behav. 1, 321-324.
- Turvey, M. T. (1967) Repetition and the preperceptual information store. J. Exp. Psychol. 74, 289-293.
- Turvey, M. T. (1972) Some aspects of selective readout from iconic storage. Haskins Laboratories Status Report on Speech Research SR-29/30, 1-14.
- Turvey, M. T. (1973) On peripheral and central processes in vision: Inferences from an information processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.
- Turvey, M. T. (1975) Perspectives in vision: Conception or perception? In Reading, Perception and Language, ed. by D. Duane and M. Rawson. (Baltimore: York).
- Turvey, M. T. (1977) Contrasting orientations to the theory of visual information processing. Psychol. Rev.
- Turvey, M. T., C. F. Michaels, and D. Kewley-Port. (1974) Visual storage or visual masking? An analysis of the "Retroactive Contour Enhancement" effect. Quart. J. Exp. Psychol. 26, 72-81.
- Turvey, M. T. and S. Kravets. (1970) Retrieval from iconic memory with shape as the selection criterion. Percept. Psychophys. 8, 171-172.
- Turvey, M. T. and R. A. Weeks. (1975) Effects of proactive interference and rehearsal on the primary and secondary components of short-term retention. Quart. J. Exp. Psychol. 27, 47-62.
- Uttal, W. R. (1969) The character in the hole experiment: interaction of forward and backward masking of alphabetic character recognition by dynamic visual noise. Percept. Psychophys. 6, 177-181.
- Uttal, W. R. (1975) An Autocorrelation Theory of Form Detection. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Von Wright, J. M. (1968) Selection in immediate memory. Quart. J. Exp. Psychol. 20, 62-68.
- Von Wright, J. M. (1970) On selection in visual immediate memory. Acta Psychol. 33, 280-292.
- Walsh, D. A. and R. E. Till. (1975) Age differences in peripheral and central perceptual processes. Paper presented at National Meeting of Gerontology Society, Louisville, Kentucky.
- Warren, R. (1975) The perception of egomotion. Unpublished Ph.D. thesis, Cornell University.
- Waugh, N. C. and D. A. Norman. (1965) Primary memory. Psychol. Rev. 72,



89-104.

Wolford, G. and S. Hollingsworth. (1974) Evidence that short-term memory is not the limiting factor in the tachistoscopic full-report procedure. Mem. Cog. 2, 768-800.

## Contrasting Orientations to the Theory of Visual Information Processing\*

Michael T. Turvey†

### ABSTRACT

In this paper the concepts of iconic memory and schematic memory are used to examine two fundamental and related features of the contemporary theory of visual information-processing. One is the orientation of indirect realism that, to put it bluntly, emphasizes the equivocality and inadequacy of the light at the eyes and the necessity of epistemic mediation. The other is the analysis of visual processing into discrete temporal cross-sections perpendicular to the flow of optical information. That the two features are closely cognate is revealed in the interpretation of event perception--the perception of change wrought over an object or object complex--as a deduction from or assimilation of (epistemic mediators) a sequence of static arrangements (discrete temporal cross-sections) represented iconically or schematically. On rational and empirical grounds, it is argued (a) that the discrete sampling of a continuous optical flow is not a tenable assumption; (b) that the informational support for event perception cannot be static iconic or schematic memories, and (c) that the perception of events cannot be epistemically mediated. Insofar as indirect realism receives little support from the analysis of event perception, direct realism is given due consideration as an alternative and radically different orientation for the theory of visual information processing.

---

\*In press, Psychological Review.

Acknowledgement: The writing of this paper was supported, in part, by the following grants to Haskins Laboratories: Grant HD-01994 from the National Institute of Child Health and Human Development and General Research Support Grant RR-5596 from the National Institutes of Health. The author wishes to express his appreciation to Carol Fowler, Robert Shaw, Robert Remez and William Mace for their conversations, questions, criticism and suggestions during the development of this paper.

†Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

## INTRODUCTION

A relatively commonplace understanding is that visual processing can be characterized as a succession of stages--of both storage and transformation--that map the arrangement of light at the receptors onto progressively more abstract representations. There are, of course, variants on this theme, mostly along lines that liberate the relations among the representations. While the more familiar hierarchy tends to define the relation between representations as unidirectional and immutable, a heterarchy permits considerable crosstalk and commutability of roles.

We should emphasize, however, that although a number of different schemes of visual information processing can be described, and although some are considerably more formal and detailed than others, it is fair to claim that they are all tokens of the same type. It is also fair to claim that, in bold outline, the current conceptions of the "flow of visual information" depart relatively little from their predecessors. Helmholtz's (1925) sketch of the "flow of visual information" in the mapping from proximal stimulus to percept, and its distant cognate, that of Alhazen a millenium before (see Lombardo, 1973), have been filled in to a significant degree by contemporary investigators and theorists, though not significantly altered. By way of illustrating this claim, here are the cornerstones of Helmholtz's theory: a finite set of primitive elements; procedures to make infinite use of this finite set; the dependence of current perception on stored memories about the world; the need for nonlocal processing to make local interpretations unambiguous; and the nonimmediacy of perception, that is, perception as a process over time (cf. Hochberg, 1974; Turvey, in press a).

The traditional and contemporary departure point for theories of visual processing is the assumption that the light at an eye is equivocal and impoverished with respect to environmental facts and events. Insofar as perception tends to be veridical, the proximal stimulus is said to underdetermine perception, and epistemic mediators are said to restore the balance. The term epistemic mediation embraces a large number of processes and representations, of which inference, short-term memory, problem-solving, the retrieving of information from permanent memory, and the comparing of current and past inputs are among the more popular. It follows from these few remarks that visual processing theories are based on a version of realism that bears the epithet "indirect." The term realism identifies a belief in an objective world, that can be perceptually experienced, and is detached from ourselves. The term indirect identifies the belief that our experience of that world is secondhand: between the world and our experience of it there intercedes a representation, or surrogate, of the world (see Shaw and Bransford, in press).

Let us consider two such surrogates, two links in the chain of epistemic mediation that characterizes a typical visual information-processing account of the mapping from proximal stimulus to percept, namely, iconic and schematic memory. These memories or representations are thought to be similar, as they both relate to the visual appearance of the stimulus (if the stimulus were a letter, neither representation would be a verbal recoding);

but they are adjudged to be quite dissimilar in several, nontrivial respects. First, and perhaps foremost, the icon is visible, but the schematic representation is not visible. Second, the icon is described as "literal" or "asymbolic," while the schematic memory is described as "abstract." Third, the upper limits of iconic persistence are thought to be considerably less than those of schematic persistence. Fourth, the icon is a maskable representation--its persistence or quality is markedly affected by temporally proximate visual stimuli--while the schematic representation seems to be impervious to masking. Fifth, the persistence function of the icon does not appear to be affected by the complexity of the stimulus, although that of schematic memory does. Sixth, the iconic representation seems to be tied to the original coordinates of the stimulus while the schematic representation does not. And seventh, iconic persistence is apparently indifferent to restrictions on processing capacity, while schematic memory, in its early stages, may be quite sensitive. For relevant references to and discussion on most of these points, see Turvey (in press a).

One may regard this paper as an exploration of the concepts of iconic and schematic memory and the visual information-processing philosophy of which they are a part. More precisely, this paper seeks to determine whether these concepts can play the roles nature requires of them. It will be shown both that the concepts are wanting and that there is reason to question the orientation of indirect realism that is the rationale for their existence.

#### RETINAL IMAGE THEORY: THE SIMULATIVE ASSUMPTION AND THE IMAGE AS AN ANATOMICAL OR ORDINAL ARRANGEMENT

Much of what is attributed to the icon--for example, that it is precatégorical, passive and impartial to the cognitive operations that follow in its wake--is due in very large part to the close (and unavoidable) relationship between the concept of the icon and the concept of the retinal image.

Perennially, the retinal image has been likened to a miniature replication or simulation of that part of the world at which the eye is directed. It has been understood, as a general tenet, that the simulation is imperfect, notably in that the retinal image has one less dimension than the world. The retinal image is only a picture.

The original impetus for the simulative assumption (Boring, 1942) and the source of its continued nourishment is the belief from indirect realism that perception is not actually of the world but of something happening in the perceiver (Lombardo, 1973). The homunculus of bygone days examined his personal copy of the world identified as the retinal image or a simulation of the retinal image. Thus, if the retinal image was a picture, then the simulation was a phenomenal picture. Echoing this precedent, the processor in current visual-information processing schemes does not process the world, but neither does "he" process the retinal image--he processes the icon (Neisser, 1967; Dick, 1974). If the icon is thought (implicitly or explicitly) to be a copy of the retinal image, then it makes little sense to think of the icon as something that is docile or malleable except in very uninterest-



ing ways. After all, no amount of experience or cognitive effort could be expected to modify the retinal image qua retinal image. Further, if the icon simulates the retinal image, then it too must be a bidimensional picture frozen in time--an internalized snapshot that corresponds to the retinal snapshot.

The ageless alternative to the pictorial view is the interpretation of the retinal image as a set of points corresponding to a set of nonintersecting light rays that can differ only in intensity and wavelength. Analogous to the pictorial view, pointillism portrays the image as bidimensional; in addition, it denies the variable of pattern.

Pointillism, of course, sidesteps the simulative assumption with respect to the relation of the retinal image to the environment, but it need not do so with respect to the relation of the retinal image to the initial internal representation. Recall that Helmholtz, among many others, viewed the initial internal representation as a set of point sensations varying in hue and brightness. In contemporary pointillism, however, reasonably sophisticated processes are said to relate the retinal image and the initial internal representation, the icon. Thus "preattentive processes" map the points into an array of segregated entities (Neisser, 1967), or feature-detection processes map the points into feature sets (Haber, 1971). Nevertheless, it is fair to claim that conceptions motivated by a pointillistic view of the icon share the paraphernalia of the simulative assumption.

Of further and even greater significance to the theory of the icon, are the following two conceptions: the retinal image as an anatomical arrangement, and the retinal image as an ordinal arrangement. (cf. Gibson, 1950). The question here, in essence, is whether the light at an eye is to be described in the coordinates of the retina or in the coordinates of the environment.

Within the anatomical conception of the image, the position and movements of a point or of a pattern are with reference to the mosaic of receptor elements, that is, to the anatomy of the retina. The pattern of light at the eye is thought of as an arrangement of receptors, and in this view, the image is said to move relative to the retina.

Within the ordinal conception, position and movement are defined relative to the arrangement of stimulation itself. One way of construing the ordinal image is as an arrangement of excitations indifferent to the actual set of receptors excited. A generalization of this construal that takes us out of the eye is more useful. The ordinal image is an arrangement of differences in intensity in different directions from the eye (cf. Gibson, 1966). This ordered discontinuity of the light at the eye is a sample of the structure of ambient light, that is, radiant light as modulated by an environment. In this perspective, the ordinal image is actually a sample of a much larger image, the ambient optic array (Gibson, 1966) in which the observer is immersed, which he scans and through which he moves. From the anatomical-image view, we say that when the eye moves, the image moves relative to the retina; in the perspective of the ordinal view, we say rather

that the retina moves relative to the image.

The distinction between these two conceptions, anatomical and ordinal, is not immaterial to visual information-processing theory. As I shall attempt to illustrate, the particular theory of processing that one constructs is profoundly determined by one's choice of image. Let us begin with a simple and fairly prosaic example. Consider a stationary point of light in an object-cluttered environment and a simple translational movement of the eyes. The point of light will change its coordinates in the retinal mosaic, and from an anatomical-image point of view, this change could be interpreted as movement of the point. How do we conclude, veridically, that it was the eyes that moved and not the point? The frequently voiced answer takes this general form: the stationary status of the point is perceived by registering the direction and extent of eye movement (information that could be provided by muscle kinesthesia or by knowledge of the efferent commands to the eye muscles, depending on one's predilections) and subtracting the computed movement from, or comparing the computed movement with, the anatomically relative movement of the point. A generalization of this answer leads to the time-honored claim that a description of a point in the coordinates of the environment is computed, with the help of extra visual information, from a description of the point in the coordinates of the eye.

The same situation--stationary point, moving eyes--receives a quite different interpretation in the ordinal-image perspective. During movement, the point's relation to the arrangement as a whole remains unaltered; in short, nonmovement of the point is specified by an invariant arrangement in the ordinal pattern. Given an information-processing device that operates on the ordinal arrangement, recourse to nonvisual information (that is, muscle-related information) and unconscious calculations of the kind described would be superfluous. Additionally, there is the implication that registering the point in environmental coordinates is not a processing step that follows subsequent to the step of registering the point in anatomical coordinates. This is not to imply that only one kind of registration occurs; indeed, both may occur and probably do, but the implication is that they are not related.

The above example is a weak illustration of the distinction between the two conceptions. It constitutes a limiting case of what there is to be processed--a point--and a limiting case of observer motion--simple eye movement. As we step up the complexity and the naturalness of the situation, the comparison sharpens.

Movements of the eyes are only one of a multitude of body activities that result in a change in the light at the retina. For example, walking, running, jumping, and rotating or flexing the head and trunk, singly or in combination, yield changes in the light at an eye that are inordinately more complex than those consequent to a simple eye movement: the latter results in a rigid translation, the former in asymmetric, topological distortions of exceptional mathematical richness. To elaborate on the situation described above, consider again a stationary point, but now permit the observer to locomote and to rotate his head. How might we describe the processing, from an anatomical-image view, that determines point stability during such a

transform? Again, we may suppose that it would be necessary to determine the resultant motion of the eyeball--of the retina--from the particular body motions responsible for the point's journey within the retinal mosaic. Then, as before, this computed eye motion could be entered into some equation with the anatomically relative movement of the point to resolve the question of whether the point moved in addition to the observer. It goes without saying that this computation of eye motion would be considerably more difficult than that for simple eye movement in a stationary head on a stationary body.<sup>1</sup> The difficulties augment quickly when we consider not a stationary point, but a stationary object, that would be projected as a particular anatomical arrangement.

Envision an object located on the ground plane and yourself as an observer walking in its direction. As you move toward and over the object, you will experience a succession of different anatomical arrangements as a result of the changing projection of that object onto the retina. Again, how might you determine whether this change is due simply to your movement and not to a movement of the object in addition to your own? The computation of the resultant eye motion will not be sufficient for this purpose. Not only will the object-produced pattern be displaced vertically from a lower to a higher region of the retinal mosaic--suggestive movement--but an increase or decrease in the number of involved receptor units is to be expected, owing to dilation and compression in the object's image, as is a change in the actual anatomical pattern owing to perspectival changes. To determine whether these variations are simply the result of your own motions, you will have to determine the shape of the object and then access knowledge about how your rectilinear motion alters the object's projected anatomical arrangement and, in addition, knowledge about how the possible motions of the object alter the projected anatomical arrangement. This solution is extremely cumbersome and appears to presuppose too much; obviously, it presupposes knowledge about how the projected anatomical arrangement of any object is altered by any of its possible motions and by any of yours.

A different and far less presumptuous solution to the foregoing puzzle derives from the ordinal conception of the light at an eye. A global transformation of the ordinal arrangement--an expanding optical flow--results

---

<sup>1</sup>Let me anticipate the developing argument; to determine the resultant motion of the eye I must have unequivocal information about the movement of the eye relative to the head, the movement of the head relative to the body, and the movement of the body relative to the environment. That is, for the formulation to work, there must be an unambiguous source of proprioceptive and exproprioceptive (see Lee, in press) information, one that does not need reinterpretation or verification by another source. Since "motor commands" and "motor effects" relate equivocally (see Bernstein, 1967; Greene, 1972; Turvey, in press b), efference signals and their copies could hardly provide such a source other than in the most trivial of cases; and since articular (joint, muscle) proprioception and vestibular exproprioception are calibrated by vision (Lee in press), they would comprise a second-rate candidate, at best, for the role of unambiguous source. But if, by implication, vision is the best candidate, then a serious paradox exists in the current formulation.

from the observer's forward, rectilinear motion and is specific to it (Lishman and Lee, 1973; Warren, 1976); in contrast, a motion of the object results in a specific transformation that is restricted to an isolated part of the total arrangement (Gibson, 1966). Thus, if both you and the object are moving, an ordered discontinuity will exist between the ordinal arrangement as a whole and a region within it. If the expansion rate of the isolated region is greater than that of the entire optical flow, then the object is moving toward you as you move toward it; similarly, if the isolated region is a "contracting" pattern within the total optical flow pattern, then you are moving toward the object but losing ground, and so on. In sum, for any particular relation between you and the object, there ought to be a specific mathematical discontinuity in the ordinal arrangement of the light at the eye. An information-processing device that is capable of registering the optical velocity field and its derivative properties will have at its disposal an optical basis for distinguishing, unequivocally, its own motions from the motions of objects in the environment that surrounds it. Formal steps toward verifying this thesis have already been taken (for example, Lee, 1974; Nakayama and Loomis, 1973).

There are many examples that could be used--for a theory of processing--to illuminate the consequences of choosing one conception of the retinal image over another. It is hoped that for our present purposes the example chosen will suffice. In the following section I shall seek to demonstrate, in part, that visual experience is sensitive to both the anatomical pattern of cells that are excited (the anatomical image) and the relational pattern of places occupied by excited cells (the ordinal image). I shall move toward the claim that experiential correlates of the anatomical pattern are rare and insignificant to visual processing in natural circumstances.

#### THE ICON IN RELATION TO THE ANATOMICAL AND ORDINAL ARRANGEMENTS

In terms of the anatomical image, position and movement are in reference to the coordinates of the eye, while in terms of the ordinal image, they are in reference to the coordinates of the environment. Afterimages are an obvious example of localization with reference to retinal coordinates: when the eyes move, an afterimage is not displaced across the retina but remains retinally fixed. A critical question is whether the icon, like an afterimage, is registered in the coordinates of the retina or whether, in contradistinction, it is registered in the coordinates of the environment. Let us adopt the oversimplification that masking is an index of the icon and proceed to respond to this question from the perspective of two ingenious experiments.

Often the claim is made that masking is very much a matter of target-to-mask proximity in the anatomical image. However, with the commonplace tachistoscopic exposure, anatomical and ordinal image are confounded. The ordinal arrangement is instantiated by the anatomical arrangement. To separate the two kinds of "images," one needs to break the coincidence of retinal and environmental coordinates. This can be done through the introduction of an eye motion between presentation of target and presentation of mask. One of the experiments to be considered examines masking with saccadic



movement (Davidson, Fox and Dick, 1973), the other examines masking with pursuit movement (White, 1976). We will consider the saccadic case first.

A horizontal array of five letters is exposed briefly just prior to a saccadic movement that carries the eyes a distance of two letters to the left of fixation. Immediately upon termination of eye movement, a mask is presented in the environmental coordinates of the fifth letter but, owing to the eye movement, in the retinal coordinates of the third letter. The query of interest is which letter will be masked. The answer is provocative: the perception of the third letter is impeded although the mask is never seen spatially coincident with the letter it masks; the mask is seen to occur in its proper environmental position and, further, appears to cover one letter while occluding the perception of another. The implication is that there is a correlate of stimulation that moves when the eyes move and a correlate of stimulation that does not. The former correlate must be of the anatomical pattern, while the latter correlate, one may venture to claim, is of the ordinal pattern. May we therefore take this experiment as evidence of a sensitivity to both arrangements and, further, as evidence to suggest that these two sensitivities might be concurrent? In this respect, note that if one has an afterimage, casting the eyes about causes the afterimage to vary its location with reference to a perceptually stable environment.

The above observation entices one to argue as follows: the correlate of the anatomical pattern is the most susceptible to masking, and since we have identified the icon as the maskable correlate of stimulation, then we can conclude that the icon is a record of the anatomical pattern of cells that are excited and, ipso facto, is localized with reference to the coordinates of the eye. By this argument, the source of iconic persistence emerges as a patch of excitation in the anatomical pattern, that is, excitation that is fixed in the mosaic of receptor cells and not displaced within the eye when the eye moves. This thesis is buttressed by an experiment with a rod monochromat which strongly suggested that the source of iconic persistence was localized primarily in the (rod) photoreceptors (Sakitt, 1975). White target letters were briefly superimposed on an intensely illuminated white background such that no matter how bright the target letters, they could not be seen. However, when the rod monochromat closed her eyes shortly after the target exposure, a visible and persisting record of the letters was experienced.

In Sakitt's interpretation, the background field saturates the rods so that any increments in intensity, such as would result from the target exposure, are indistinguishable. Since the letters are eventually distinguishable, the source of their persistence must be localized prior to the first stage of the visual system that saturates. The weight of the evidence suggests that the first stage that saturates is most likely the rod photocurrent and the source of the visible icon is adjudged, therefore, to be in the photoreceptors. When the rod monochromat closes her eyes, the rods start to recover and the photocurrents dip below their saturation level. With the advent of this state, those photoreceptor locations that were more strongly stimulated--that is--those coincident with the letter display, induce a larger neural signal than the surrounding photoreceptor locations

that were stimulated only by the background field. It is this "larger neural signal" that is said to give rise to the visible icon subsequent to closing of the eyes.

One might inquire whether this conclusion, that the source of iconic persistence is retinal--is contradicted by the fact that iconic persistence can be altered by dichoptically opposed stimulation (for example, Meyer, Lawson, and Cohen, 1975). We can answer confidently that it is not. The persistence in the firing of photoreceptors means that neuroanatomical channels allied to the receptor surface will be active for comparable periods. However, if neuroanatomically more central components of these channels are perturbed in some fashion--such as by cross-adaptation--they may not be able to match the persistence of photoreceptor activity, and the visual experience, for which they are partly responsible, will be affected similarly.

Curiously, the conclusion that the icon is the correlate of the anatomical image is not supported by the second of the two experiments that examine masking during pursuit movement. For when a target and an aftercoming mask are delivered to an observer who is visually tracking a moving dot, masking is maximal when the target and mask share the same environmental coordinates, and minimal when they share the same retinal coordinates (White, 1976). Mimicking the line of reasoning used above, one might now be compelled to argue that the icon is a correlate of the ordinal arrangement.

To summarize, if we choose to define as iconic that correlate of stimulation that is maskable, then there is not one style of iconic persistence but two--one that is in the coordinates of the retina and one that is in the coordinates of the environment. Official doctrine informs us that a description of a stimulus in the coordinates of the environment is mediated by a description in retinal coordinates. This is an assertion of the primacy of the anatomical arrangement in visual processing, and in part, the remarks in the previous section were meant to illuminate this sentiment. Now, given the two contrasting "icons," we should suppose, therefore, that within the official doctrine, the icon in environmental coordinates is computed from that in retinal coordinates. The experiments of Davidson et al (1973) and White (in press) offer little support for the orthodox view; indeed, they are more suggestive of a dissociation rather than a dependency. In the pursuit-movement experiment (White, 1976), masking in environmental coordinates occurred at target-to-mask intervals of 50 and 100 msec intervals at which masking retinal coordinates was absent. Unfortunately, mask delays of less than 50 msec were not assayed, but the masking functions reported suggest that an even finer time titration would not insure masking in retinal coordinates, but would continue to insure masking in environmental coordinates. If true, this extrapolation is paradoxical given the traditional thesis that an object's environmental position is computed over time (with the aid of eye-movement information) from the object's position in the retinal mosaic.

Tentatively, we may hypothesize that the correlate of stimulation in environmental coordinates does not depend on that in retinal coordinates,

although the two may occasionally interact. In the context of the previous section, the hypothesis takes this form: processing the anatomical arrangement and processing the ordinal arrangement are largely separate activities.

Let us take a further, guarded step. Let us conjecture that in visual processing au naturel the ordinal arrangement is what is processed, and the anatomical arrangement is ignored. Under certain conditions an experience of the anatomical arrangement may obtrude on the experience of the ordinal arrangement, but such obtrusion is rare. Generally, the anatomical image is transparent (in the literal sense of the word) to the information that is processed.

"Transparency" is the key to understanding masking in retinal coordinates. A saccadic movement per se does not erase the anatomical pattern of a previous fixation (Doerflein and Dick, 1974). Erasure results from the new anatomical arrangement that is consequent to the saccade, superimposed on the old. With saccadic movements, the successive anatomical arrangements are discrete and separate from each other. On the other hand, the ordinal arrangement subsequent to a saccade is neither entirely new nor separate from its predecessor, since any two successive samples have a large overlap, the maximum displacement of an eye never exceeding its angular field. Indeed, almost all the fixations in the field of view of the head will have something in common with their topological neighbors. The degree of overlap is the degree of structure common to the successive samples; the ordinal arrangement that remains unchanged during movements of the eyes. Doubtless, what remains unchanged corresponds to the scene the observer is examining (Gibson, 1966). At all events, we may recognize that the masking in retinal coordinates, the annihilation of an anatomical arrangement that accompanies a saccade, enhances in simple but elegant fashion the transparency of the anatomical image. This would be to the benefit of a perceptual system that processes the mathematical order of stimulation, ideally unencumbered by its anatomical manifestation.

In some such way as the preceding argument, we may rationalize the retinally-related masking that is incident to saccadic movement. How is the environmentally related masking incident to visual tracking to be understood? White (in press) has remarked that since one's "goal" in pursuit movement is to maintain a fixation on an object, the perceptual suppression of nontarget, background objects is salutary. Thus in pursuit movement, masking is functionally more useful in environmental than in retinal coordinates. However, it is very difficult to conceive such masking ever occurring in more natural circumstances. Abrupt discontinuous changes at an environmental site are most improbable, and evolution would not have considered it an occurrence worthy of special mechanisms. It would seem that the rationale for environmentally related masking will have to be found elsewhere.

#### THE DISCRETE, STATIC SAMPLE ASSUMPTION: ICONIC AND SCHEMATIC STORAGE IN RELATION TO EVENT PERCEPTION

I have referred to both the correlate of the anatomical arrangement and the correlate of the ordinal arrangement as iconic storage. At this time, it

is both advisable and reasonable to identify the correlate of the ordinal arrangement as schematic memory and thus avoid an unnecessary proliferation of icons. Although it was suggested earlier that schematic memory is essentially nonmaskable, and therefore unlike the icon, I choose now to qualify this assertion by remarking that schematic memory may be maskable under certain conditions--namely, pursuit movement. There is much to encourage this definitional alteration and consequent relabeling; most notably the congruence of the overall distinction between iconic and schematic memory, commented on earlier, and the distinction developed over the last few pages between anatomical and ordinal sensitivity.

However, focusing on the iconic/schematic distinction should not blind us to the shared characteristic of the two concepts: both are conceived as static "snapshots." This commonality bears significantly on the question I shall pursue: given the concepts of iconic and schematic storage, how might we account for the perception of events, that is to say, the perception of change wrought over an object or complex of objects?

To set the stage for the critical issues that follow, consider a relatively mundane event such as the rotation of an object. Obviously, rotation occurs over time and, equally obvious, the optical information needed to determine that the object is rotating, is temporally extensive. Let us refer to this flow of optical information as the transforming optic array (Gibson, 1966) and contrast it with the "information flow" with which we have been concerned in visual information-processing theory, that might be described as the flow of "neural information" or the transforming "neural array" (Neisser, 1972).

Herein lies, as far as one can discern, the most significant feature of visual information processing as a philosophy and methodology: the analysis of visual perception (or cognition) into discrete temporal cross-sections perpendicular to the flow of optical information. Essentially, the thrust of research and theory has been to describe the vicissitudes of retinal input, that is, the sequence of transformations that the retinally related neural array undergoes, within a discrete temporal cross-section. The in theory argument is that a combination of these discrete temporal cross sections will yield the registration of an event, for which the optical information is realized over time. Considerable lip service has been paid to the eminent plausibility of this in theory argument, but curiously, the content of the argument has rarely, if ever, been examined. Furthermore, the enormously influential assumption of discrete sampling--and we may remark that it is, indeed, the primitive of information-processing theory--has undergone little serious scrutiny.

Wherein lies the motivation for the discrete sampling assumption? Once again we may identify the culprit as the retinal image construed as an anatomical pattern. On this point, a further historical detour will serve us well.

At the onset of the 17th century, Kepler had been convinced by Della Porta of the appropriateness of the camera obscura as a metaphor for the eye,



a metaphor that had been proposed originally by Alhazan and reinforced by Da Vinci (Lombardo, 1973). This conviction paved the way for Kepler's explanation of how the light rays are focused to produce a two-dimensional image on the surface of the retina. The camera metaphor had served a useful didactic purpose; however, it may be argued that the metaphor it was to shackle thinking on visual processing in ways that, in retrospect, far outweighed its original usefulness. To pay heed to the camera metaphor is to liken the retinal photoreceptors to photographic film. Consequently, it must be argued that in order to avoid blurring, the retinal "film" must be exposed to incident light for but a brief moment, just enough to permit the capture of an image. Contemporary theorists, of course, decry the eye-as-a-camera theme; nevertheless, as far as the conventions of visual processing theory are concerned, the die had been cast many centuries ago. The eye had been characterized as a device for capturing static images, and the visual system as an instrument for analyzing them. The perception of form or static pattern was defined as the basic area of investigation, and the perception of change, of events, was conceptualized as a deduction from sequences of static arrangements.

A commonplace current rationalization for the discrete-sampling assumption stems from the fact that perceiving an object-cluttered environment is achieved via scanning, where scanning is interpreted as a succession of discrete, static arrangements each separate from its neighbor. Respectfully, however, we should note that the static character of the "snapshot" provided by a fixation is guaranteed only if the observer is stationary and is scanning an unchanging, temporally frozen environment. The absence of movement in observer and environment is a limiting case, and we will have to inquire, as before, what happens when one considers other, more natural instances of observer/environment interaction. First, however, let me preface these considerations with a perfunctory analysis of the limiting case.

Each fixation in a series of fixations results in a new anatomical image. Given that the dwell time--the duration of a fixation--in natural scanning is of the order of 200 msec, it has seemed most reasonable (to some students of perception) to contend that the perception of an unchanging scene is mediated by a succession of temporal cross sections of iconic size (cf. Haber, 1971). Is perceiving a scene, then, just a matter of adding together persisting icons? Most obviously it is not, for as we have come to understand, each new anatomical arrangement annihilates the previous one; it follows that this must be the fate of successive icons as correlates of anatomical arrangements. Consequently, we are led to propose that whatever the function operating on the discrete temporal cross section, its variable is more likely to be the schematic rather than the iconic form of memory. In brief, we can think of a schematic memory as derived, during a fixation, from iconic memory or from the fixed ordinal arrangement; in either case, we can then speak of schematic "snapshots" and conjecture that a function on these determines the perception of a static scene.

Let us at this point unfreeze the environment and the observer. Consider situations in which one is scanning a dynamically transforming scene while locomoting or observing a rotating object from a fixed position of the

eye. Here again, traditional predilections might invite an analysis of these occurrences into successions of "snapshots" (cf. Neisser, 1967). Unfortunately, there is no obvious or simple mechanism for decomposing the optical flow into a series of discrete patterns such as that provided by successive fixations. Does not a dynamically transforming scene require such decomposition within fixations in order to differentiate the optical flow into static patterns? This appears to be a thorny problem for the theory of visual information processing. With a continuously changing optical flow pattern, there is a continuously changing anatomical and ordinal arrangement. And in the absence of a source of discontinuity other than fixations, it would be nonsensical to speak of discrete anatomical or ordinal arrangements and their respective iconic and schematic correlates as the informational support for the perception of events.

However, it does not stretch the imagination to concoct other sources of discontinuity. Here are two. First, discrete samples, corresponding to discrete anatomical arrangements, could be achieved by a mechanism that repeatedly turns the retinal mosaic on (for very brief periods) and off (for comparatively long periods). For didactic reasons let us entertain such a mechanism, however, improbable it may be. Second, discrete static patterns could be achieved by a mechanism that sampled, discontinuously, the optical flow further upstream in the neural flow--precisely--at the level of schematic memory.

Considering a mechanism that could package the continuous optical flow for processing into successive, temporally discrete samples, let us make critical inquiry into the style of a system that takes discrete inputs as intrinsic to its operation.

A time-honored distinction between perception and memory takes this form: the domain of perception is adjacent order, as given in a simultaneous composite of elements; the domain of memory is successive order, as given in a temporally distributed collection of elements. Consequently, to register an event involves, among other things, perceiving an adjacent arrangement and storing its percept, perceiving the next adjacent arrangement and storing its percept, and so on. From this perspective, to register an event requires both perception and memory. For purposes of illustration let us translate "percept" into "schematic memory" and let us identify the storage medium into which schematic memories are entered as short-term memory qua a medium that preserves short-term memories. Here, then, is a common formula for explaining the registration of events that occur over reasonably short periods of time; a succession of snapshots is entered into short-term memory and then analyzed or integrated in some fashion. An inquiry into the role adduced for short-term memory in this formula and also into the form of the proposed analytic or integrative operation, will prove instructive.

Consider Figure 1. Decomposition of the optical flow into discrete patterns is assumed at the retinal level, but it could just as well be assumed at the schematic level. The successive snapshots (anatomical arrangements) of the events are ordered on the time line T. The internal representations (schematic memories) of these snapshots are ordered in short-

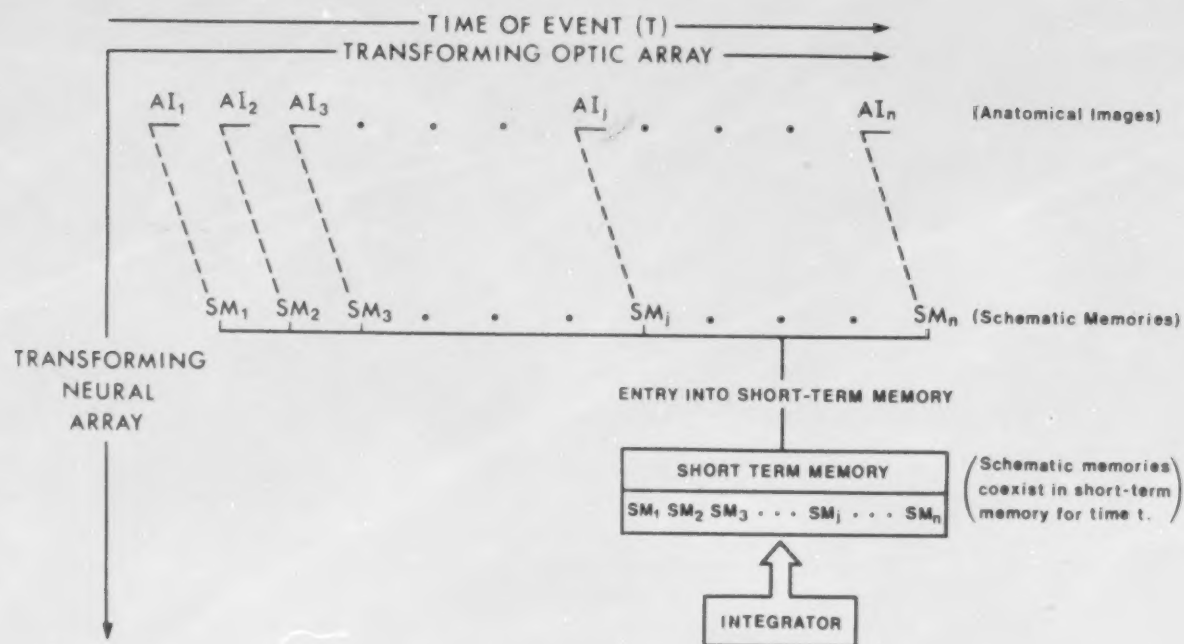


Figure 1: The contrasting predictions of the Discrete Moment and Traveling Moment Hypotheses (after Allport, 1968).

term memory on the time line  $t$ , where  $t < T$ . On inspection of Figure 1, one is struck by the realization that the role attributed to short-term memory is that of compressing the time over which the internal representation of the event is distributed. Ideally, the mechanism of short-term memory makes the successive discrete samples contemporaneous. In its ideal form this role of short-term memory manifests a most subtle alchemy: the conversion of successive arrangements to an adjacent arrangement, the translation of time into space. If short-term memory mimics such an alchemist, then presumably, perception--which, according to the time-honored distinction, is sensitivity to adjacent order--can take over to detect the event that transpired! Unfortunately, in this account, we have gone from perception to memory and back again to perception, and the problem of explaining how an event is registered given a successive ordering of adjacent arrangements now becomes the problem of how an event is registered given only adjacent order. Of course, we could assume that the schematic memories are tagged for time-of-entry into short-term memory, but then we would have to suppose some internal process, akin to perception or to inference, that scans the memories according to their temporal order. Needless to say, that is much like the problem we began with, namely, explaining how an event--a change over time--is registered; and it suggests that we have gained little in the way of explanatory yardage from our appeal to the concept of short-term memory.<sup>2</sup>

#### Schematic Maps and Frame Systems

There is a perfectly good reason for rejecting the notion that the successive snapshots are preserved in short-term memory preparatory to registering an event: more often than not, the number of snapshots would exceed the useful capacity of the memory (Hochberg, 1968; 1970). Therefore, instead of a process that, say, draws inferences from a set of snapshots (schematic memories) in storage, we might wish to propose a set of operations that construct the event from the successive glimpses as they occur, and thereby ignore short-term memory altogether. This has been the route championed by Neisser (1967), Hochberg (1968) and Minsky (1975), although

---

<sup>2</sup>Consider the problem of "slow events," changes that occur over long periods of time (for example, aging). To explain our apprehension of such things we might wish to make an appeal to the concept of long-term memory that is understood (similarly) as a collection of schematic memories. These memories, we may suppose, have been garnered erratically during the unfolding of the event. If cognizance of a slow event is interpreted as an inference from a circumscribed collection of schematic memories, then as a preliminary step, we must explain how any one schematic memory of the event comes to be related in long-term memory to any other schematic memory of the event. What relates two or more (indeed, all) of these schematic memories is that they are instances of the same object under transformation. To establish the identity relation between the schematic memories necessitates knowledge of the transformation (see Shaw and Pittenger, in press). In sum, to relate the schematic memories in a way that makes possible the inference that this or that event transpired, presupposes knowledge of the event that transpired.



only the latter two have provided anything like a formal account of such a procedure.

The essential idea is this: when one encounters a situation, one selects from one's long-term model of the world a substantial structure called a schematic map (Hochberg, 1968) or a frame (Minsky, 1975) that will be adapted to fit the situation by changing details as necessary or discarded if a fit cannot be made, in which case another map or frame is then selected. Given the right map or frame, the successive snapshots are fitted together, or encoded into the remembered structure. It is this assimilation that determines the perception of a scene or situation. Hochberg defines a schematic map in this fashion: "The program of possible samplings of an extended scene and of contingent expectancies of what will be seen as a result of these samples" (Hochberg, 1968:323). In brief, a schematic map is a matrix of space-time expectancies (Hochberg, 1968). A frame is defined similarly (Minsky, 1975). It is a data-structure for representing a stereotyped situation; it is said to contain information about how to use the frame, what can be expected to happen next, and what to do if the expectations are not confirmed. Actually, when speaking of events, we should be considering frame systems rather than frames, where the different frames of a system capture the object or scene from different perspectives, and where change of scene or object is represented within the system by the transformation from one frame to another (Minsky, 1975). And, in like fashion, if we were to address events from Hochberg's stance, we should have to include what he calls schematic sequences, in addition to the schematic maps already described. Crudely speaking, schematic sequences are remembered temporal orders.

Enough has been said to suggest convergence between the terms proposed by Minsky and Hochberg, so further reference to schematic maps and sequences is not needed. We can take refuge in the more neutral and less abused terms of frames and frame-systems.

A terse deliberation must suffice to illustrate what is entailed by a frame-systems approach to event perception. At the outset, we should suppose that context supplies a suitably colored backdrop of evidence and expectation for determining the selection of an initial frame system; otherwise, as with any other analysis-by-synthesis scheme, the process of registering the event would proceed. Yet we must appreciate that events, like words, can be perceived with facility in the absence of, and out of, context and, that the perception of context itself must rely on frame systems.

What if the initially selected frame system provides less than an adequate fit to the event? Presumably there must be strategies which guarantee that the next system chosen provides a better fit. When there is a misfit, there must be a procedure that distinguishes between the object and the form of change (the two aspects of the event instantiated by the frame system) as the source of incongruence. Furthermore, in the course of verifying a selected frame system, that is, while constructing the event, a record must be kept of those snapshots of the event that have been "absorbed" already; otherwise, they may be reanalyzed and included as new information

(Hochberg, 1970). This latter requirement sounds very much like an appeal to a short-term memory for successive snapshots, a mechanism I had hoped to avoid.

These preliminary remarks on a frame-systems approach are testimony to the inelegant and perhaps, implausible device that perceives events by reference to prior conceptual knowledge about events. Let us pass, therefore, from frame systems to a further consideration of the discrete sampling assumption, and more in the direction of hypothesis that asserts, contrary to tradition, that the registration of events is not predicted on the registration of static patterns incident to discrete sampling.

#### Discrete Moment Versus Traveling Moment

Of immediate interest is the distinction between the discrete moment hypothesis and the traveling moment hypothesis. A simple metaphor, due to Allport (1968), contrasts the two hypotheses rather nicely. Consider a person standing on a railway platform looking into the windows of a passing passenger train for an expected friend. For our observer, the passengers on the train are revealed compartment by compartment as each window passes his or her point of view. In this sense, the glimpses our person has of the interior of the train are essentially discontinuous in time. They are analogous to a series of snapshots in which the presence of a feature or object in one snapshot excludes its presence in the next.

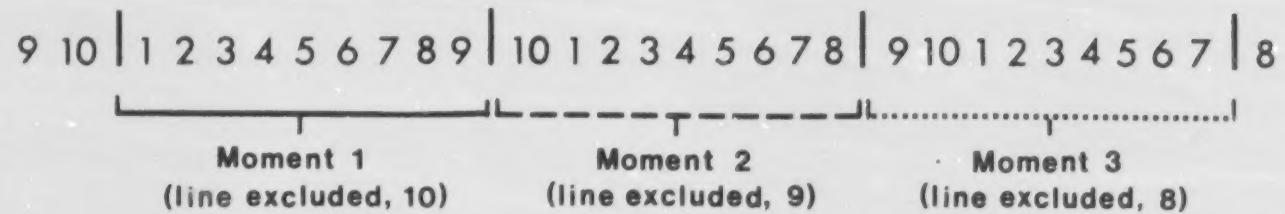
Suppose now that we consider the scenario from the perspective of the expected friend inside the train. The field of view that he or she has of the platform is bounded by the compartment window; new aspects of the platform are revealed continuously at one side of the window, while other, older aspects are occluded continuously at the other side of the window. The moving window is our metaphor for "traveling moment."

There is another experiment also due to Allport (1968), that puts the two hypotheses in competition, and from which the traveling moment hypothesis emerges as victor. Suppose that one presents ten lines in rapid succession on a cathode-ray tube, beginning by displaying the line at the bottom of the tube, then turning it off and displaying the line just above it, then turning it off, and so on. The procedure continues until the 10th and top line has been displayed and turned off, at which point the procedure starts again with the bottom line. One can adjust the rate of presentation so that all lines are visible simultaneously; a slight reduction of this rate will result in nine lines visible simultaneously, with one missing. The missing lines will appear as a shadow that moves gradually across the face of the tube. Now, interestingly, the discrete moment hypothesis predicts that the "shadow" will appear to travel in the direction opposite to that of the sequence of line positions, while the traveling moment hypothesis predicts motion of the shadow in the same direction as the sequence of line positions. The two predictions are depicted in Figure 2. It should be remarked, by way of summary, that observers unanimously see the shadow move in the direction predicted by the traveling moment hypothesis. It thus appears that processing is continuous, not intermittent.

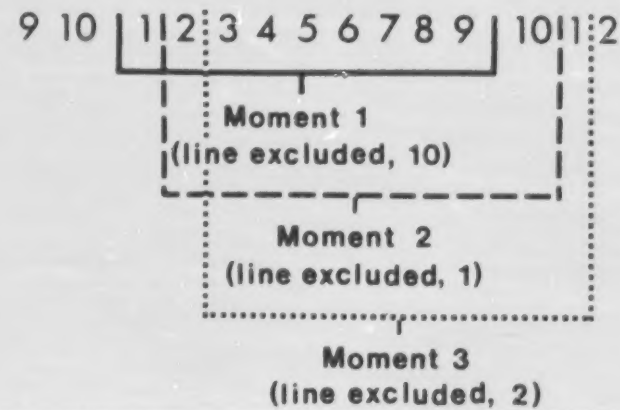
Figure 2:

In one traditional view, the continuous optical flow concomitant to an event is decomposed into a sequence of static arrangements--anatomical images--which in the form of schematic memories are brought temporally closer together, even to the point of contemporaneity, by the function of short-term memory.

### Discrete Moment Hypothesis



### Travelling Moment Hypothesis



There is yet another very remarkable phenomenon that militates against the discrete sampling notion.<sup>3</sup> When a cube is rotated on a face, its period of symmetry is four, since every 90 degrees of rotation brings the cube into self congruence; for rotation on a corner and rotation on an edge, the periods of symmetry are three and two, respectively. If a wire cube rotating at constant speed is strobed, a procedure that simulates discrete "snapshots", then whether or not an observer sees a cube and/or rotation depends on whether or not the strobe frequency is an integral multiple of the extant period of symmetry (Shaw, McIntyre, and Mace, 1974). By inference, if in everyday circumstances the perceiver samples at an immutable rate, then the nonveridical perception of events should be commonplace; for surely, only on rare occasions would the perceiver's sampling rate be synchronized with the symmetry period of the event. Being able to control the rate at which the optical flow corresponding to an event is discretely sampled would not help either, because, for veridical perception, the perceiver would have to know in advance the nature of the object, the nature of the change, and the period of symmetry. To emphasize the absurdity of the discrete sampling notion, consider the problem of perceiving under natural circumstances two cubes rotating simultaneously, one on a face and one on an edge. In this case, in order to see the two events, the observer simultaneously would have to sample simultaneously at two different rates.

Let me complete this brief survey of phenomena unfavorable to the discrete sampling assumption with the following observation: patients with visual agnosia syndromes sometimes manifest a peculiar dissociation between (a) static bidimensional patterns and static three-dimensional objects and (b) these same patterns and objects when either they or the patterns and objects are moving (Zappia, Enoch, Stamper, Winkelman, and Gay, 1971; Botez, 1975). The patients are "blind" to the static objects, but they recognize them under dynamic transformation. If, again, we were to suppose that visual processing proceeds from static samples--procured by a stroboscopiclike operation--then, not unreasonably, we would have to expect these patients to be visually agnostic under any condition of observation, static or dynamic.

In these last two phenomena, the rotating cube and the agnosia, we can dimly see the outline of a potentially significant premise for the theory of visual information processing: structural variation is not inferred from structural nonvariation, but rather, structural variation reveals structural nonvariation.

#### Limitations of Iconic and Schematic Memory

Our inquiry has brought us, in short, to the conclusion that discrete sampling of a continuous optical flow is not a tenable hypothesis. The concepts of iconic and schematic representation, as they are most generally construed, are viable only in the context of discontinuous processing. Most

---

<sup>3</sup>I am indebted to Carol Fowler for pointing out the relevance of this phenomenon to the discrete versus continuous issue.



obviously, a reappraisal of these concepts and of the visual information-processing theory to which they are allied, is in order.

Most significantly, neither the icon nor schematic memory can be proposed as the object of perception in the sense of providing a data base for visual information processing. Generally, neither iconic nor schematic memory can be said to mediate perception or to be intrinsic to it. These representations have assumed significant roles in our theories because we have limited our experimental analysis to the perception of a frozen scene at a single glance from a stationary point of observation. That which we have labeled iconic storage is essentially the record of the anatomical pattern of excited cells, a record which is only isolable and persistent under very special constraints on viewing--for example, tachistoscopic exposure. There is no reason to believe that anything comparable to iconic persistence is manifest when one views a scene undergoing change or when one is in motion. If iconic persistence is potentially present in the circumstance of scanning a frozen environment from a stationary point of observation, there is no reason to believe that it contributes to the perception of the scene attendant to scanning. What then are we to make of the icon and its properties? The answer might be that the icon and its properties index some aspects of the neuroanatomic structures that support the act of perceiving; they are not, however, indices of processes constituent of that act.

Consider the following metaphor. I wish to understand the workings of a telescope--how it magnifies--but my investigation draws me to chromatic aberrations manifest by one of the lenses. Using available techniques, I monitor this characteristic and probe the rules that govern its occurrence. The balance of the metaphor leaves little to the imagination: the object I have chosen to study is a property of the machinery that supports magnifying, but it is hardly intrinsic to the process.

But what of schematic memory? It has been defined as a correlate of the ordinal image, which is defined in turn as the relational pattern of places occupied by excited cells. The ordinal image incident to a fixation, however, is a limiting case of ordinal arrangement. When the eyes scan a frozen scene, successive fixations as purveyors of ordinal pattern are neither separate nor isolated; on the contrary, they are overlapping and mathematically related. In short, they comprise among themselves an ordinal arrangement. Similarly, and more prominently, in the case of a dynamically transforming environment and/or a moving observer, there is, over time, both change and nonchange in ordinal structure, more precisely, in the ambient optic array. Consequently, if schematic memory is defined as the correlate of the ordinal arrangement, then the static representation or ordinal structure incident to a fixation is simply the limiting case of schematic memory. And insofar as our acquaintance with schematic memory is restricted to the limiting case, we may concur that schematic memory is not something we know much about. We can claim, however, the schematic memory is not a constituent of, but a consequence of, visual processing--a consequence of the pickup of invariant and variant structure in the flux of light at the eye.

A REAPPRAISAL OF THE VISUAL PROCESSING ASSUMPTIONS:  
THE PERSPECTIVE OF DIRECT REALISM

Let us conclude with a critical examination of the information processing enterprise. As we have seen, visual information processing, as a methodology and as a philosophy, is devoted to the solution of the following problem: how is the information in the light entering the ocular system mapped onto perceptual experience?

On the topic of problem solving, Simon (1969) comments that it is essentially a matter of translating, back and forth, between two descriptions of the same phenomenon. On the one hand, there is the state description, that characterizes the phenomenon as we sense or intuit it, and on the other hand, there is the process description, which characterizes the phenomenon in terms of how we might produce it. As Simon remarks: "The general paradigm is; given a blueprint, to find the corresponding recipe." (p. 112).

A Consideration of the State Description for Haptic Information Processing

The state description of the visual information processing problem was identified at the outset, namely, that visual perceptual experience commonly corresponds veridically to the properties of the environment even though the proximal stimulus relates imperfectly to those properties. We have said, of the proximal stimulus, that it underdetermines perception. Consequently, the search is for that recipe, or at least that class of recipes, that will effect this blueprint, the blueprint of indirect realism. Suppose that we are mistaken in our description--that we have the wrong blueprint--then it matters little what recipes we discover.

The point of departure for these concluding remarks is the conjecture that the state description of the visual information-processing problem has been inadequately specified, with nontrivial consequences for the understanding of the "how" of processing. Indeed, we will acknowledge the point of view that the most significant determinant of the manner in which we construe perceptual processing to occur is what we conceive the information in the proximal stimulation to be (cf. (Gibson, 1966; Garner, 1974)). In a phrase, our conception of the what of processing determines our investigation and interpretation of the how of processing. Let us return to our earlier discussion of Helmholtz.

Recall that in response to the hypothesized equivocality of proximal stimulation, Helmholtz proposed that a man or woman perceives that particular state of the environment that would normally fit the particular proximal stimulus received. For Helmholtz, knowledge of what is normal, or most probable, is given through experience. According to the Helmholtz position, the sensory mechanisms responding to the light at the eyes are not especially elaborate, as they yield only a patchwork of colors. The task of the perceiver is to recover from this colored but relatively formless tapestry the state of the environment. What the perceiver must learn, therefore, are procedures that permit him or her to do just that. How can one acquire such procedures when one's visual relationship with the world is so limited? The

visual sensations (or features), after all, are uninformative about the facts of the environment. Evidently, knowledge about the facts of the world must be culled from some source other than vision. The tradition of empiricism has looked to mechanical commerce with the world as the means by which the impressions of vision are rationalized: proximal distributions of light come to be associated with particular tactile/muscle-kinesthetic states.

For example, I adjust my hands, that is, tailor the arrangement of the joints of the fingers, to the shapes of objects. However, the immobile clasp of the hands provides a far from perfect mold even for those objects small enough to be grasped, and we do, after all, perceive objects in great detail, whether the objects be large or small. Supposedly, the tactile information about shape comes not from immobile clasps but from dynamic exploration. Similarly, there is the matter of locomoting toward objects and the fact that the amount of effort expended relates in some fashion to the distance traversed. This is perhaps, in part, how one comes to interpret the distance of objects from the size of their retinal images, as Berkeley (1871) argued. Against this notion, however, are many counterpoints. To take but one, the retinal image size that is contiguous, and thus supposedly associated with the degree of kinesthesia and effort expended in locomoting, is that retinal image size which occurs at the end of locomotion and not that which was present at the outset. So how can kinesthetic correlates of the distances traversed to reach objects provide a basis for inferring distance from retinal-image size? This particular formula would work only if in our earliest years, we spent most of our time crawling and walking backwards! Undaunted, one could point out that the significance of mechanical commerce follows from the fact that it is intrinsically reinforcing. A hot stove or a lighted match is directly painful to the possessor of the hand that touches it. In this role, tactile stimulation would tell us only that we liked or disliked some particular optical arrangement; it could not in any obvious fashion specify, say, the shape of the distal object to which the particular optical arrangement corresponded.

One might wish to argue that these last remarks do not necessarily allay the claim that tactually exploring and manipulating the environment is the source of knowledge about its properties. For the percept of an object involves more than an impression of its shape and its distance from us; it includes a detailed knowledge of the object's properties such as solidity, stickiness, sharpness, heaviness, coarseness, and so on. We have many intuitive reasons for describing these as nonoptical properties. As Gregory (1969) remarks, "To build a seeing machine we must provide more than an 'eye' and a computer. It must have limbs, or the equivalent, to discover nonoptical properties of objects for its eye's images to take on significance in terms of objects and not merely patterns" (p. 246). The hypothesis, most obviously, is that mechanical commerce with the environment is privileged in that it permits the direct detection of object-properties. In short, mechanical commerce with the world is directly meaningful. This is a most curious point of view, for it eschews recognition of the fundamental sensationalism of the Helmholtz-type theory. Is it not the case that when an object deforms the skin, or induces a changing configuration of joint postures, all that is directly available is an array of tactile/kinesthetic

sensations or features? (Although as far as I know, features have never been identified for haptic processes.) Is it not true that this tactile/kinesthetic "patchwork" demands interpretation as much as its visual counterpart? We may legitimately assert that the traditional empirical solution to the problems raised by the nonspecificity and piecemeal nature of the proximal stimulus in vision is no solution at all. This is so because faithful adherence to the traditional perspective results in assigning meaning to meaningless optical patterns by correlating them with meaningless tactile/kinesthetic patterns. Most obviously, the empirical account of tactile/kinesthetic perception ought to look just like the empirical account of visual perception; in which case neither perception, in theory, ought to occur.

Nevertheless, it is instructive to be occasionally delinquent in one's faith. If tactile/kinesthetic processes do register object properties in a fashion that does not have to resort to epistemic mediators, that is, to sources of knowledge other than the stimulation itself--a fashion which might be referred to as "direct"--then the ambiguity of the visual proximal stimulus is not a recalcitrant problem for an experience based theory of the Helmholtz kind, though it remains a difficult one. So let us, for pedagogical purposes, grant to tactile/kinesthetic processes this special virtue. Consequently, for mechanical commerce with the environment, the state description of the information-processing problem is as follows: tactile/kinesthetic or (more appropriately) haptic perceptual experience corresponds to the properties of the environment because those properties determine the pattern of haptic stimulation, which in turn determines the perceptual experience. Let us seek a corresponding process description for this state description.

Mechanical commerce with the environment permits the discovery of object properties, or so we hypothesize. Suppose that I wish to determine the consistency of an object; I will squeeze, prod, punch, or stretch the object. If I wish to assess its texture, I will run my fingers--or some suitable extension of my fingers, such as a stick--over its surface; for, interestingly enough, I cannot obtain reliable discriminative information about surface texture by mere contact or pressure (Meenes and Zigler, 1923; Katz, 1925). To determine the mass of an object so that I might compare it with the mass of another, my best policy is to wield it in various ways. Throwing or tossing the object is far more informative about its mass than is passive grasping (Gibson, 1966). From such observations, we infer the significance of change, that is, of transformations in tactile deformation and patterns of joint articulation, to the determination of object properties.

Obviously, since haptic perception is not epistemically mediated, ex hypothesi, there has to be, for any object property, some component of the changing stimulation which remains unchanged and which corresponds unequivocally to the property in question. How should we characterize this invariant? It is, in principle, an often complex mathematical arrangement that is revealed in the course of change. For each property of objects and surfaces detectable by the haptic system, we must suppose that there is a unique mathematical arrangement defined across the concomitant patterns of



skin contacts and joint articulations covarying in time.

Roughness presents a limited "instance." The perception of roughness is only slightly affected by a 25-fold increase in the rate at which the fingers move across a surface, and relations among roughness perceptions of different surfaces are unaffected by the force with which the fingers are pressed against them (Lederman, 1974). A possible (quasistatic) candidate for the relevant invariant is the cross-sectional area of the amount by which the skin is depressed below its overall level (Taylor and Lederman, 1975).

At all events, by collecting the immediately preceding paragraphs, we may claim that haptic perception depends on the detection of information about something, in the sense of specificity to something, that is preserved over exploratory transformation. Thus, one part of the process description that we seek is clearly defined although difficult to implement--the delimiting and manipulating of invariants under relevant transformations. Another part follows.

In pursuing the gist of Helmholtz theory, we might propose that there is a set of elementary properties, each property specified by a particular haptic arrangement, from which all other properties are derived. The distinguishing feature of these elementary properties is that they are the direct givens, in the sense that the processes supporting their detection or determination, do not have to resort to sources of knowledge other than the stimulation itself. No epistemic operations (for example, matching with memory or crosscorrelating with the data of other modalities) are said to intercede between stimulation and the registering of these elements; the criteria for their detection are wholly in the stimulation (cf. Uttal, 1975).

If the elementary properties are a, b, and c, what shall we say of property d? We can say that it is derived from a, b, and c and that its detection is by virtue of a special process, that is, one different in kind from that by which a, b, and c are detected. Since d is not an elementary property, then by definition, the criteria for its detection are not wholly in the stimulation, but reside somewhere else in memory. Obviously, the special process of which we speak is an instance of epistemic mediation. However, epistemic mediation in haptic perception has been ruled out, ex hypothesi, so the detection of d cannot entail a special process that is qualitatively different from that supporting the detection of a, b, and c. Following this line of reasoning, we reach a most curious conclusion: there can be no partitioning of haptic stimulation into primary and directly given properties on the one hand, and secondary and derived properties on the other. Invariants specifying environmental properties may lie on a continuum from less to more complex, and they may be detectable with different degrees of facility, but the manner of their detection must be the same--that is, direct. In sum, the process description corresponding to the aforementioned state description of haptic perception is charged with specifying an account of perceptual processing that asserts the primacy of abstract, relational information and the direct pick-up of such information. It cannot operate on the nominal principle that the perception of some set of elementary proper-

ties is a necessary prerequisite to the perception of other, higher order, properties.

Taking stock of our analysis of haptic processing, it is evident that we have been describing a version of realism that is more aptly termed direct, in contrast to the indirect realism that customarily provides the backdrop for the inquiry into visual processing. By direct realism we understand, in part, that it is the world that is actually perceived and not some surrogate of it (see Shaw and Bransford, in press).

This account of haptic information-processing as nonmediated was necessitated by the desire to salvage the empiristic formulation of visual information-processing. That is, if haptic processing could reveal object properties unequivocally, then these products of haptic perception could be associated in experience with visual patterns and thus give the latter meaning. However, if we drop the stance of empiricism and accept phylogenetic experience as a viable source of knowledge about the world, then perhaps the above description of haptic perception is unwarranted. The significant question is this: can we imagine a set of information-processing systems, each operating in the Helmholtz style, uncovering in the course of evolution those worldly facts and laws necessary to the interpretation of ambiguous and piecemeal sense data? That is to say, can the mediators of perception possibly have a perceptual origin? If they cannot, then we are confronted by an a priori statement of the worst kind: for any species, knowledge about the world was present from the very beginning. This kind of a priorism would seem to characterize the approach to scene analysis by machine (see Sutherland, 1973). In reference to the question of how to program the perceptual process into a computer, Hunt (1975) remarks that we can do so only by finding a way to represent the world in the machine. In short, for the machine to come to perceive its world, it must already know its world.

One may hazard a guess that in order for an animal designed Helmholtz-style, to evolve and relate adaptively to its ecological niche, it would be necessary for at least one perceptual system to acknowledge environmental properties without epistemic mediation. We could thus argue from a phylogenetic perspective, much as we argued from an ontogenetic perspective, for the direct realism of mechanical commerce with, or haptic processing of, the environment. We should ask: if evolution chose to go the way of direct realism with one processing system, why would it not choose to go that way with the others?

The main conjecture at which we have now arrived, in regard to the theory of visual information processing, is that we may look upon direct (but, of course, critical<sup>4</sup>) realism as the departure point for the theory. It is in the light of this conjecture that we take one further glimpse at the perception of events.

---

<sup>4</sup>In naive realism, it is assumed that all things can be perceived, in critical realism we assume that not all things can be perceived. Direct but critical realism, means that what can be perceived is perceived directly.

### The Pursuit of Invariants Defined Over Continuous Optical Transformations

In our previous considerations of event perception, we learned of the nominalistic tenor of orthodox visual-processing theory. Precisely, with an event defined as the change wrought over an object or arrangement of objects, the "information" processed relevant to the resultant experience of the event is that of static bidimensional forms or static perspectives. Accordingly, the experience of the style of change is an abstraction from elementary aspects. What the orthodox view denies, is the possibility of abstract relations--defined on optical structure over time--that are specific to the style of change per se. Indubitably, it is the preclusion of such abstract entities, invariants, that entices--even demands--the interpretation of event perception as a synthesis, either from the memories of static perspectives and intellectual inference, or from the assimilation of static perspectives by one of a large number of stored frame systems. It is exactly such abstract entities and their apprehension without reference to memories and concepts, or any other form of epistemic mediation that is championed in the direct-realism approach to visual processing (Gibson, 1966; Turvey, 1975).

There is an instructive illustration to be found in the simple change undergone by an object in its motion from one location to another. In an indirect-realism view, this event--translation--is perceived by virtue of the prior discrimination of spatial and temporal extent. The motion parameters of velocity and acceleration are inferred by noting the positional changes of the object and the time over which the changes occur; for acceleration, the spatial and temporal extents between two successive positions would need to be computed on several separate occasions in order for there to be any degree of accuracy. The reader will agree that for this formulation to work, it is not sufficient to take discrete samples; rather, the discrete sampling must occur at a constant rate that must be known. Inasmuch as we have previously concluded that such discrete sampling is implausible, it would be odd if the evidence were favorable to this formulation of indirect realism. Fortunately, it is not. It can be shown that the perceptions of velocity and acceleration are not determined by the prior discrimination of discrete spatial and temporal positions, but instead appear to be directly perceived attributes of the translatory event (Lappin, Bell, Harm, and Kottas, 1975; Rosenbaum, 1975). In keeping with the expectations inherent in a direct-realism view, the evidence is for abstract information defined on the relation between spatial and temporal changes that is specific to these parameters of translation (see Lappin et al, 1975). The corollary is no less important, namely, that spatial and temporal positions are not perceptual primitives from which are composed the perceptions of velocity and acceleration.

In our reflections on haptic processing as nonmediated, there was a conclusion reached to which we may now return and recognize as a fundamental precept of direct realism. The perception of each and every event or property is based on invariant information detected over time, and not on the discrimination of elementary aspects (Gibson, 1966). As previously remarked, the two most significant components of events are the nature or style of the change (for example, rotating, bouncing, running) and the object properties

preserved over the change. It follows, therefore, that for direct realism there must be two kinds of invariant information in the ambient optic array corresponding to these two components--invariants that might be labeled, respectively, transformational and structural (Pittenger and Shaw, 1975). A transformational invariant is that information specific to style of change that is preserved over different structures "supporting" the change; a structural invariant is that information specific to object structure that is preserved over the styles of change in which the object participates.

Herein lies an important and far-reaching contrast with the story of event perception as told in the language of indirect realism. For in that language, there is no equivalent to transformational invariant, and the concept closest but not identical to a structural invariant, is the static silhouette of an object as given by discrete sampling (cf. Pittenger and Shaw, 1975). In an indirect realist's account of event perception, knowledge about the style of change must be presupposed; it resides in memory by sleight of hand. In a direct realist's account, however, information about the kind of change is postulated to exist unequivocally in the transforming optic array. We may say of transformational and structural invariants that they are "formless" (Gibson, 1973), for the crux of the matter is that no "form" remains in a continuous transformation--"form" or "perspective" or "silhouette" is annihilated. What does remain are those aspects of forms or of styles of change that are invariant over time. It is in this sense that invariants are said to be formless, for they are not themselves forms. The summary of these remarks has a curious ring to it: for the direct realist, the perception of events depends on the detection of formless invariants and not on the perception of forms.

Our simplified reconnaissance of direct visual processing carries us a little way toward an appreciation of the contrasting perspectives for a theory of visual processing. To pursue this subject further would take us too far afield for the purposes of the present paper. Others have sought to sharpen and clarify the issues and distinctions I have barely touched upon, and to these authors the reader should turn for a fuller account (Gibson, 1966; Mace, 1974, in press; Shaw and McIntyre, 1974; Shaw and Bransford, in press; Turvey, 1975).

There is one aspect of the contrast on which we should comment by way of conclusion. Presumably, the goal of visual processing theory is to isolate and characterize that which is most eminently and directly responsible for our perceptual knowledge. In the view of indirect realism, the candidates for this honor are the postulated links in the internal chain of epistemic mediators from retinal image to perceptual experience. However, the view of direct realism promotes a very different roster of candidates. They are, most obviously, the complex nested relationships in the dynamically structured medium surrounding the observer that are specific to the properties of the environment in which he or she acts. What is prescribed by this latter view is an ecological attitude to research and theory on visual processing; an "ecologizing" of physics, as it were, so that we might better comprehend how the world relates to humans and animals as knowing agents rather than as physical or biological objects (Shaw and Bransford, in press). An ecological



attitude suggests a different tack in research from that which is currently being navigated as well as a different batch of skills from those at the disposal of contemporary explorers of visual information processing.

In sum, the issue of whether perception is direct or indirect cannot be treated lightly, for the consequences of its resolution will be far from trivial in the quest for an understanding of how we know our world by sight.

#### REFERENCES

- Allport, D. A. (1968) Phenomenal simultaneity and the perceptual moment hypothesis. Brit. J. Psychol. 59, 395-406.
- Berkeley, G. (1871) An essay toward a new theory of vision. (1709) In The Works of George Berkeley 1, ed. by A. C. Fraser. (Oxford: Clarendon Press).
- Bernstein, N. (1967) The Co-Ordination and Regulation of Movements. (New York: Pergamon).
- Boring, E. G. (1942) Sensation and Perception in the History of Experimental Psychology. (New York: Appleton Century Crofts).
- Botez, M. I. (1975) Two visual systems in clinical neurology: Readaptive role of the primitive system in visual agnostic patients. Europ. Neurol. 13, 101-122.
- Davidson, M. L., M. J. Fox, and A. O. Dick. (1973) Effects of eye movements on backward masking and perceived location. Percept. Psychophys. 14, 110-116.
- Dick, A. O. (1974) Iconic memory and its relation to perceptual processing and other memory mechanisms. Percept. Psychophys. 16, 575-596.
- Doerflein, R. S. and A. O. Dick. (1974) Eye movements and perceived location in iconic memory. Paper presented at Psychonomic Society Meetings, Boston.
- Garner, W. R. (1974) The Processing of Information and Structure. (Hillsdale, Md: Lawrence Erlbaum Assoc.).
- Gibson, J. J. (1950) The Perception of the Visual World. (Boston: Houghton Mifflin Co.).
- Gibson, J. J. (1966) The Senses Considered as Perceptual Systems. (Boston: Houghton Mifflin Co.).
- Gibson, J. J. (1973) On the concept of "formless invariants" in visual perception. Leonardo 6, 43-45.
- Greene, P. H. (1972) Problems of organization of motor systems. In Progress in Theoretical Biology, vol. 2, ed. by R. Rosen and F. M. Snell. (New York: Academic Press).
- Gregory, R. L. (1969) On how little information controls so much behavior. In Toward a Theoretical Biology, vol. 2, ed. by C. H. Waddington. (Chicago: Aldine Publishing Co.).
- Haber, R. N. (1971) Where are the visions in visual perception. In Imagery, ed. by S. Segal. (New York: Academic Press).
- Helmholtz, H. von. (1925) Treatise on Psychological Optics, ed. and trans. from the 3rd German ed. (1909-1911) by J. P. Southall. (Rochester, N.Y.: Optical Society of America).
- Hochberg, J. (1968) In the mind's eye. In Contemporary Theory and Research in Visual Perception, ed. by R. N. Haber. (New York: Holt, Rinehart

- and Winston).
- Hochberg, J. (1970) Attention, organization and consciousness. In Attention: Contemporary Theory and Analysis, ed. by D. I. Mostofsky. (New York: Appleton-Century-Crofts).
- Hochberg, J. (1974) Higher-order stimuli and inter-response coupling in the perception of the visual world. In Perception: Essays in Honor of J. J. Gibson, ed. by R. B. MacLeod and H. L. Pick, Jr. (Ithaca: Cornell University Press).
- Hunt, E. (1975) Artificial Intelligence. (New York: Academic Press).
- Katz, D. (1925) Der Aufbau der Tastwelt. (Leipzig: Barth).
- Lappin, J. S., H. H. Bell, O. J. Harm, and B. Kottas. (1975) On the relation between time and space in the visual discrimination of velocity. J. Exp. Psychol.: Human Percept. Perform. 1, 383-394.
- Lederman, S. J. (1974) Tactile roughness of grooved surfaces: The touching process and effects of macro- and microsurface structure. Percept. Psychophys. 16, 385-395.
- Lee, D. N. (1974) Visual information during locomotion. In Perception: Essays in Honor of James J. Gibson, ed. by R. B. MacLeod and H. L. Pick, Jr. (Ithaca, N.Y.: Cornell University Press).
- Lee, D. N. (in press) On the functions of vision. In Modes of Perceiving and Processing Information, ed. by H. Pick and E. Saltzman. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Lishman, J. R. and D. N. Lee. (1973) The autonomy of visual kinaesthesia. Percept. 2, 287-294.
- Lombardo, T. (1973) J. J. Gibson's ecological approach to visual perception: its historical context and development. (Unpublished Ph.D. Thesis, University of Minnesota).
- Mace, W. M. (1974) Ecologically stimulating cognitive psychology: Gibsonian perspectives. In Cognition and the Symbolic Processes, ed. by W. Weimar and D. S. Palermo. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Mace, W. M. (in press) James Gibson's strategy for perceiving: Ask not what's inside your head, but what your head's inside of. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. E. Shaw and J. Bransford. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Meenes, M. and M. J. Zigler. (1923) An experimental study of the perceptions roughness and smoothness. Am. J. Psychol. 34, 542-549.
- Meyer, G. E., R. Lawson, and W. Cohen. (1975) The effects of orientation-specific adaptation on the duration of short-term visual storage. Vision Res. 15, 569-572.
- Minsky, M. (1975) A framework for representing knowledge. In The Psychology of Computer Vision, ed. by P. H. Winston. (New York: McGraw Hill).
- Nakayama, K. and J. M. Loomis. (1973) Optical velocity patterns, velocity-sensitive neurons and space perception: a hypothesis. Percept. 3, 63-80.
- Neisser, U. (1967) Cognitive Psychology. (New York: Appleton Century Crofts).
- Neisser, U. (1972) On "Going beyond the information given": A reply to J. J. Gibson's memo. Unpublished manuscript. (Cornell University).
- Pittenger, J. B. and R. E. Shaw. (1975) Aging faces as viscal-elastic events: Implications for a theory of non-rigid shape perception. J. Exp. Psychol.: Human Percept. Perform. 1, 374-382.

- Rosenbaum, D. A. (1975) Perception and extrapolation of velocity and acceleration. J. Exp. Psychol.: Human Percept. Perform. 1, 395-403.
- Sakitt, B. (1975) Locus of short-term visual storage. Science 190, 1318-1319.
- Shaw, R. E. and J. Bransford. (in press) Introduction: Psychological approaches to the problem of knowledge. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. E. Shaw and J. Bransford. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Shaw, R. E. and M. McIntyre. (1974) Algoristic foundations to cognitive psychology. In Cognition and the Symbolic Processes, ed. by W. Weimer and D. S. Palermo. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Shaw, R. E. and J. B. Pittenger. (in press) On perceiving change. In Modes of Perceiving and Processing Information, ed. by H. Pick and E. Saltzman. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Shaw, R. E., M. McIntyre, and W. Mace. (1974) The role of symmetry in event perception. In Perception: Essays in Honor of J. J. Gibson, ed. by R. B. MacLeod and H. L. Pick. (Ithaca, N.Y.: Cornell University Press).
- Simon, H. (1969) The Sciences of the Artificial. (Cambridge: MIT Press).
- Sutherland, N. S. (1973) Intelligent picture processing. Paper presented at Conference on the Evolution of the Nervous System and Behavior, Florida State University, Tallahassee.
- Taylor, M. M. and S. L. Lederman. (1975) Tactile roughness of grooved surfaces: A model and the effect of friction. Percept. Psychophys. 17, 23-36.
- Turvey, M. T. (1975) Perspectives in vision: Conception or perception? In Reading, Perception and Language, ed. by D. Duane and M. Rawson. (Baltimore, Md.: York).
- Turvey, M. T. (in press a) Visual processing and short-term memory. In Handbook of Learning and Cognitive Processes, V, ed. by W. K. Estes. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Turvey, M. T. (in press b) Preliminaries to a theory of action with reference to vision. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Uttal, W. R. (1975) An Autocorrelation Theory of Form Detection. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Warren, R. (1976) The perception of egomotion. (Unpublished Ph.D. thesis, Cornell University).
- White, E. W. (in press) Visual masking during pursuit eye movements. J. Exper. Psychol.: Human Percept. Perform.
- Zappia, R. J., M. Enoch, R. Stamper, J. Z. Winkelman, and A. M. Gay. (1971) The Riddoch phenomenon revealed in non-occipital lobe lesions. Brit. J. Ophthalmol. 55, 416-421.

Evidence for a Special Speech-Perceiving Subsystem in the Human\*

Alvin M. Liberman+ and David B. Pisoni++

ABSTRACT

If we want to discover whether man is specialized to process speech so as to recover phonetic segments, we must, of course, make the appropriate comparisons with nonhuman animals. To promote that undertaking, we here identify a distinctive characteristic of phonetic (as opposed to auditory) perception, and we describe some phenomena of human speech perception, appropriate for testing with animals, in which that characteristic seems to be exhibited. The distinctive characteristic is that the perceptual process is constrained as if by 'knowledge' of what vocal tracts do when they make linguistically significant gestures. The distinctive phenomena are taken from instances of stop-consonant perception. There, the role of a necessary acoustic cue--silence--is to inform the listener that the speaker closed his vocal tract, as he must if he is to produce a stop consonant; and the equivalence in perception of very different acoustic cues--temporal vs. spectral, for example--is to be accounted for on the ground that, though presumably unrelated in auditory perception, they are the distributed results of the same articulatory gesture.

If it were possible to perceive the words of language simply as auditory patterns--that is, without regard to their constituent phonetic elements--then neither phonetic structure nor its perception would be of great biological interest. But such a nonphonetic strategy would, in practice, severely limit the number of words a listener could identify and immensely complicate the processes by which he extracts those words from the stream of speech.

Our assignment is to ask whether biologically special processes might be necessary for perceiving the phonetic structure of language. It seems appropriate, then, to consider those facts about speech and its perception

---

\*This paper was presented at the Dahlem Konferenzen, Berlin, October, 1976, and will be published in the conference proceedings Dahlem Konferenzen, ed. by T. H. Bullock.

+Also University of Connecticut, Storrs.

++Indiana University, Bloomington, Indiana.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]



that imply the need for such processes, and to sharpen the point by imagining experiments with nonhuman animals and human infants aimed at finding out whether they hear speech as we do.

We will distinguish two kinds of processes among those that might be specialized for speech perception. One includes specializations of the auditory system that would serve to extract from the complex speech signal just those parts that carry the linguistically relevant information. We should wonder whether such devices exist if only because there appears to be a need for them: paradoxically, some of the acoustic cues that underlie important linguistic distinctions are among those aspects of the physical signal that are least salient. If there are specializations of that kind, they might be similar to the feature detectors that have been claimed for so many animals. At all events, they would properly belong to the auditory system, however specialized for speech they might be, because they would succeed only in clarifying the signal; they would not decode it. There would remain a peculiar relation between auditory pattern and phonetic message, a relation similar in form and function to those grammatical codes that connect other levels of language in the further reaches of phonology and syntax. Conceivably, there are devices specialized to cope with that peculiar relation, and thus to recover the phonetic message from the sound. If so, they would presumably be different in kind from the specialized auditory devices we have imagined. Indeed, such specialized devices would likely be an integral part of the larger and equally special physiology that comprehends all of the grammatical link between sound and meaning. Hence, we will distinguish specialized grammatical decoders from auditory specializations; the grammatical decoders we are concerned with in this paper would most properly be considered to be specializations of a phonetic sort.

If auditory or phonetic specializations for language do exist, we should expect that, given the appropriate experimental tests, the responses of nonhuman animals would be different from ours; the responses of human infants would, of course, depend additionally on the way those specializations are affected by experience. To promote consideration of how we might, in any case, do relevant research, we will identify several classes of findings with adult human beings that suggest what some of the animal and infant tests might be. But, given the limitations on the length of this paper, we will reserve the matter of auditory specializations for the discussions we expect to have with the members of the conference. Here we will concern ourselves only with the question: Are there specialized phonetic processes?

To provide a proper background for our question, we should remind ourselves of two universal--hence biologically interesting--facts about language. The more obvious is that the structure of language has two aspects: one is formed of meaningful segments (words, phrases, sentences) and governed by the rules of syntax; the other comprises segments that are empty of meaning (phones, syllables, breath groups) and subject to the lawful constraints of phonology and phonetics. Less obvious, but no less universal, is the fact that the shortest of the meaningless segments--the phones, or consonants and vowels, that are the objects of our attention here--are not directly reflected in the sound stream. That is so because of the universal occurrence of coarticulation, the overlapping or even simultaneous production

of features from successive phonetic segments. As a consequence, information about those several phonetic segments is carried simultaneously on the same acoustic parameter. Thus, the phonetic message is encoded (not enciphered) in the sound, and in a special way: there is no direct correspondence in segmentation between message and sound, hence no acoustic (or auditory) criterion by which the speech stream can be directly divided into segments that correspond to the phones; and the acoustic cues for any particular segment will vary, often in apparently peculiar fashion, according to the other segments with which it is encoded and simultaneously conveyed.

It is the existence of that universal (among human beings) and special code that most generally bespeaks the need for special phonetic processes. Accordingly, a decent concern for the importance of putting first things first dictates that we should want most urgently to know how well nonhuman animals cope with its most general characteristics. Can they, for example, appreciate, even tacitly, that speech does consist of commutable segments, that "bad" and "dab" are simply different permutations of the same three segments, or that words like "grew" and "ilk" share no segments but have the same number? Unfortunately, the animal tests appropriate to those most general, and possibly most telling, questions are often impossible in practice, or so nearly so as to discourage even the most intrepid investigators. With that in mind, and in the hope that relevant experiments of some kind might nevertheless be done, we will set considerations of logical priority aside and give special emphasis to those less general and more simple--yet still apparently special--characteristics of phonetic perception for which the appropriate animal tests might be feasible. And in order to crowd as many of those characteristics as we can into our allotted space, we will, to the greatest extent possible, deal with a single and simple acoustic cue, silence. (For further discussion of relevant data and issues, see: Darwin, 1976; Liberman, 1974; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Pisoni, in press; Stevens, 1975; Stevens and House, 1972; Studdert-Kennedy, 1976.)

#### Some Acoustic Cues Have Perceptual Effects That May Be Uniquely Phonetic

Consider the following easy-to-obtain facts about the way we perceive stop consonants and fricatives. Record the syllable /sa/, represented schematically in Figure 1. There you see a patch of band-limited noise, normally produced in the articulation of /s/, followed by a vocalic section. The vocalic section contains, first, the formant transitions caused by the articulatory movement from consonant position to vowel position, then the formant steady-states appropriate to a drawn-out vowel. Since the partial closure for the fricative /s/ is at approximately the same place in the vocal tract as the total closure for the stop consonant /t/, the formant transitions of /sa/ are similar to those of /ta/. We find, then, that if we remove the patch of noise, human listeners will, indeed, commonly hear not /a/, but /ta/. Restore the noise, and they will, of course, hear /sa/ again. Move the noise backwards in time so as to open up a gap, or silent interval, of about 60 msec between it and the vocalic section, and they will hear not /sa/, but /sta/.

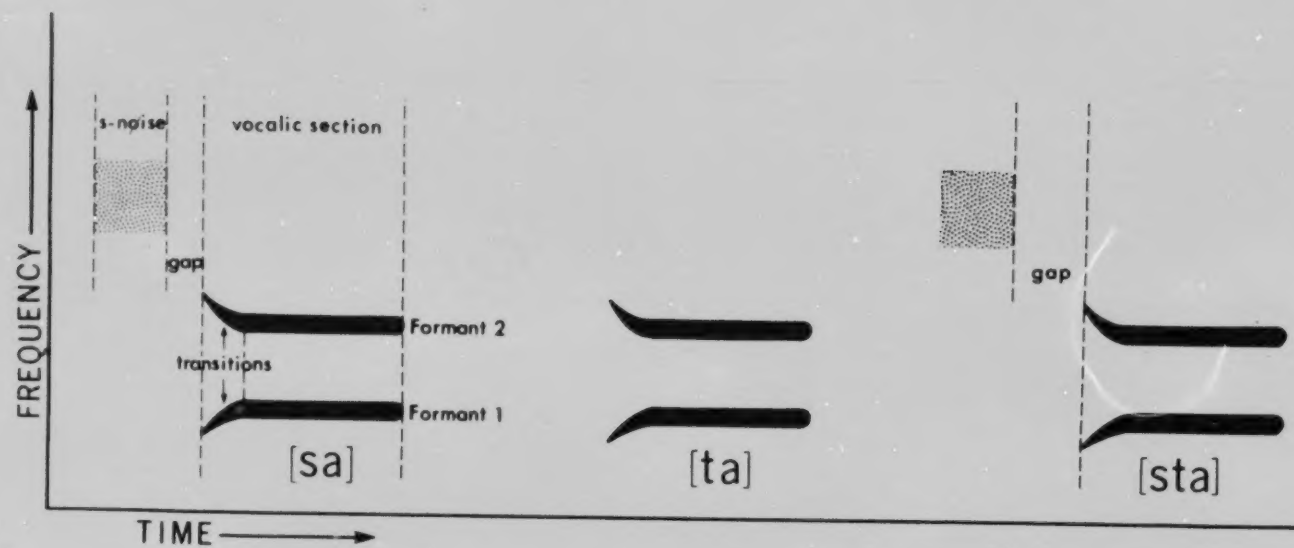


Figure 1: Schematic representation of variations on the syllable [sa], illustrating the role of silence in the perception of the stop consonant [t].

Thus, silence is a condition for perceiving the stop consonant. But what does silence do? From the point of view of our interests at this conference, there are at least three possibilities. The first is that the role of silence is explicable in terms of the properties of a generalized auditory system. Consider, once again, the phenomena described above, and see that the formant transitions, which are cues for the stop consonant, might be forward-masked by the noise; in that case, the silence would provide time to evade masking. Given that kind of explanation, we should expect that animals with ears like ours would hear the syllables much as we do.

A second possibility is that we have here the result of an auditory specialization of the kind we referred to earlier. There might, for example, be detectors specialized to extract formant transitions from speech, and these might be disabled by the noise; or there might be specialized auditory devices that produce an interaction between silence and the transition cues. In either case, other animals would not hear the syllables as we do; but in human listeners the observed effects would be found in all of auditory perception, not just in speech.

The third possibility is that the perception of silence is here phonetic rather than auditory. To distinguish phonetic from auditory, we should determine whether the perceiving mechanism is constrained, not only by the properties of the ear, but also by "knowledge", as it were, of what vocal tracts do when they make linguistically significant gestures. To see how just such a constraint might be at work in our example, consider that a speaker cannot say /sta/, as against /sa/, without totally closing his vocal tract and so creating an interval of silence. Given a biologically based link between speech perception and speech production, the absence of silence might therefore signal the listener that the speaker had not closed his vocal tract long enough to have said /sta/; hence /sa/. Such phonetic perception, if it does exist, would be found only in creatures that speak, and then only when they are listening to speech.

We will shortly describe some facts about the perception of speech by human beings that, in the case of the silence cue and others, imply the existence of a phonetic mode of perception. We will look forward then to learning as much as possible about how nonhuman animals hear these same speech sounds, and how both they and human listeners hear the relevant acoustic variations in a nonspeech context. But before we abandon the simple example of /sa-sta/ that we have already offered, we should at least list the several reasons we have for supposing that it is a simple case of phonetic perception: (1) Though perception of the stop can be totally blocked by the s-noise, the transition cues are nevertheless effective in promoting perception of the fricative /s/ (Harris, 1958; Darwin, 1971). Thus, as auditory events, the transition cues "get through";<sup>1</sup> it is only their (phonetic) interpretations (as /s/ or as /t/) that are affected by the nearness (or farness) of the noise. (2) When the noise of /s/ is put very close to the vocalic syllable /ka/, listeners hear neither /ska/ nor /sa/, but /sja/.<sup>1</sup>

---

<sup>1</sup>Lieberman, Halwes, and Fitch: personal communication.



Consider that the transition cues for the stop /k/ and the semivowel /j/ are similar except that the former are more rapid than the latter. We see, then, that the transition cues were interpreted as /ja/, not /ka/, because the gap that is so essential for the production and perception of /k/ was not there; to produce /ja/, the speaker does not close the tract totally, hence he produces no gap. Here, too, it was only the phonetic interpretation of the transitions (as /k/ or as /j/) that was affected. (3) The transition cues in a syllable perceived as /se/ are fully effective in producing selective adaptation of /de/ (Ganong, 1975). This is further evidence that, as auditory events, the transition cues are, in fact, being processed by the perceptual system; they are in no way blocked by the presence of the s-noise. (4) When inserted between s-noise and a vocalic section, silence is sufficient (and not merely necessary) for the perception of a stop consonant; in such cases, there are no ordinary stop-consonant cues (transitions or bursts) to be masked or detected (Dorman, Raphael, and Liberman, 1976). (5) In cases like those described in (4), the "place" of the perceived stop--that is, whether it is /b/ or /g/--depends on the nature of the following vocalic section.<sup>2</sup> (6) In other, analogous cases, the amount of silence necessary to produce a stop-like effect varies according to the tempo of the carrier phrase (Raphael, Dorman, and Liberman, 1976). (7) When the transition cues for the stop are removed from the vocalic speech contexts and presented alone, in which case they are heard as nonspeech "chirps," their identifiability is not at all masked by the preceding s-noise, nor is their perception changed in any qualitative way (Dorman, Raphael, Liberman, and Repp, 1975).

Different Acoustic Cues Produce the Same Phonetic Perception in the Same Position and in the Same Context

We offer the example of the contrast between /slit/ and /split/, diagrammed in Figure 2 so as to show how it can be fashioned out of either of two very different cues--one spectral, the other temporal.<sup>3</sup> The spectral cue is primarily the appropriate set of formant transitions; it is present (plus /p/) in the vocalic section /plit/ and absent (minus /p/) in /lit/. The temporal cue is the silent gap by which the vocalic sections are separated from the s-noise; it is present (plus /p/) when the gap is long and absent (minus /p/) when it is short. In Pair 1, a spectral difference is sufficient to cue the contrast between /slit/ and /split/. In Pair 2, we see how that same contrast is produced by an acoustic cue that is entirely temporal. One asks, of course, what the spectral and temporal cues have in common. In acoustic and auditory terms they appear to be about as different as can be. However, from an articulatory point of view they are related: they are the results of the same gesture. Given that they have the same consequences in phonetic perception, we might suppose that in this case perception is somehow linked to production.

<sup>2</sup>Summerfield and Bailey: personal communication.

<sup>3</sup>Liberman, Halwes, and Fitch: personal communication.

FIGURE 2

|          | DESCRIPTION OF STIMULI |                                               | PERCEPT                      | CHARACTERIZATION OF CUES |          |           |           |
|----------|------------------------|-----------------------------------------------|------------------------------|--------------------------|----------|-----------|-----------|
|          | GAP                    | VOCALIC                                       |                              | TEMPORAL                 | SPECTRAL | TEMPORAL  | SPECTRAL  |
| PAIR I   | s noise                | short --- <i>lit</i><br>short --- <i>plit</i> | <i>slit</i><br><i>split</i>  | -p<br>-p                 | -p<br>+p | same      | different |
| PAIR II  | s noise                | short --- <i>lit</i><br>long --- <i>lit</i>   | <i>slit</i><br><i>split</i>  | -p<br>+p                 | -p<br>-p | different | same      |
| PAIR III | s noise                | short --- <i>lit</i><br>long --- <i>plit</i>  | <i>slit</i><br><i>split</i>  | -p<br>+p                 | -p<br>+p | different | different |
| PAIR IV  | s noise                | short --- <i>plit</i><br>long --- <i>lit</i>  | <i>split</i><br><i>split</i> | -p<br>+p                 | +p<br>-p | different | different |

Figure 2: Diagrams that show how very different acoustic cues--one temporal, the other spectral--produce the same phonetic distinction, and how both cues together either enhance that distinction or reduce it depending on the way they are combined.

Different Acoustic Cues for the Same Phonetic Perception Cancel or Summate Depending on How They are Combined

Using the same cues described above, we find that they have opposite effects depending on how they are put together. In Pair 3, we show how to combine the spectral and temporal cues so as to produce the same phonetic distinction (/slit/ vs. /split/) that either cue alone is sufficient to make. We should note that when each cue is so near the perceived phonetic boundary as to be less than perfectly unambiguous, that combination will result in a distinction that is even more robust. In Pair 4, we see how those same two cues can be combined so as to decrease, not increase, the phonetic difference that is produced by either cue alone. In fact, that combination can, with proper balancing, effectively bring the difference near to zero. Thus, it is as if these cues were vectors; but however we might characterize them, we shall suppose, until animal tests prove us wrong, that their domain is not auditory but phonetic.

The Perceptual Effects of Acoustic Cues are Subject to Ecological Constraints of an Apparently Phonetic Kind

Acoustic cues can have one phonetic effect or another depending on whether they were produced by one speaker or by two (Dorman, Raphael, Liberman, and Repp, 1975; Raphael, Dorman, and Liberman, 1976). It is as if the listener knew that two vocal tracts can accomplish what one vocal tract cannot. To see an example, imagine the following. We record "now say" and, separately, "shop". Then we place "shop" after "now say". If the physical characteristics of the word "shop" are held within certain limits, it is possible to change "now say shop" to "now say chop" simply by increasing the duration of silence between "say" and "shop". Of course, this is just like the other examples of silence we have described; and in this case, as in the others, we assume that the silence causes the listener to hear the affricate (in "chop") rather than the fricative (in "shop") because an appropriate amount of silence tells him that the speaker closed his vocal tract briefly, as one must to produce the affricate, and as one must not to produce the fricative.

However, two vocal tracts--one saying "now say" and the other "chop"--can produce "now say chop" with no silence at all between "say" and "chop". Thus, with two speakers, the size of the interval of silence provides no useful phonetic information. Experiments reveal that listeners behave accordingly: starting with "now say" and "shop", and given a silent interval appropriate for "chop", listeners do indeed hear "now say chop" if there was only one speaker; but if there were two speakers, then listeners hear "now say shop" at all intervals of silence.

### Different Acoustic Cues Produce the Same Phonetic Perception in Different Positions

Consider the voicing distinction between, for example, /b/ and /p/ in three positions in the syllable, as schematized in Figure 3. In initial position, as in /bid/ vs. /pid/, an important and sufficient cue is the so-called voice onset time (VOT), the time interval between release and start of voicing (Liberman, Delattre, and Cooper, 1958; Lisker and Abramson, 1964). In intervocalic position, as in /rabid/ vs. /rapid/, an important acoustic cue is the duration of silence between the two syllables (Lisker, 1957). And in final position, as in /ib/ vs. /ip/, an important and sufficient cue is the duration of the preceding vowel (Raphael, 1972). We should note how very different these cues are from an acoustic point of view. We should also note that in each of these cases, as in the earlier example of /slit/ vs. /split/, there are several cues, very different from each other in acoustic terms, but equivalent in phonetic perception. That is most dramatically the case with the contrast between /rabid/ and /rapid/, where the effective acoustic cues include, in addition to the duration of intersyllable silence, such variables as the duration of the first syllable, the spectral characteristics (transitions) at the end of the first syllable and at the beginning of the second syllable, the condition of syllable stress (whether trochaic or iambic), the voice-onset-time in the second syllable, and numerous others.<sup>4</sup> As in the case of /slit/ vs. /split/, this diversity of acoustic cues is produced by a single articulatory contrast. That such acoustically different cues are more or less equivalent in phonetic perception is further testimony to the link between the way we perceive such phonetic distinctions and the way we produce them.

### The Same Acoustic Cues are Perceived Differently According to the Remote Context

Consider, again, the perceived difference in voicing between /rabid/ and /rapid/ as cued by the duration of silence. When experimental tokens of such syllables are placed in speech carriers that mimic different rates of articulation, listeners hear the change from /rabid/ to /rapid/ at different durations of silence (Port, 1976).

### Different Acoustic Cues Produce the Same Phonetic Perception in the Same Position but in Different Immediate Contexts

Unfortunately for our purposes, the silence cues do not offer telling examples in this case. We are reluctant on that account alone to omit the kind of context-conditioned variation we would here illustrate, because it is one of the most important consequences of coarticulation, hence one of the most pervasive characteristics of the speech signal. We will, then, turn away from our preoccupation with duration and the sounds of silence just long enough to present a well-worn example of what can happen when we change only one phonetic segment in a syllable that contains three. Take /did/ and /dud/, shown schematically as two-formant (synthetic) approximations in

<sup>4</sup>Lisker: personal communication.



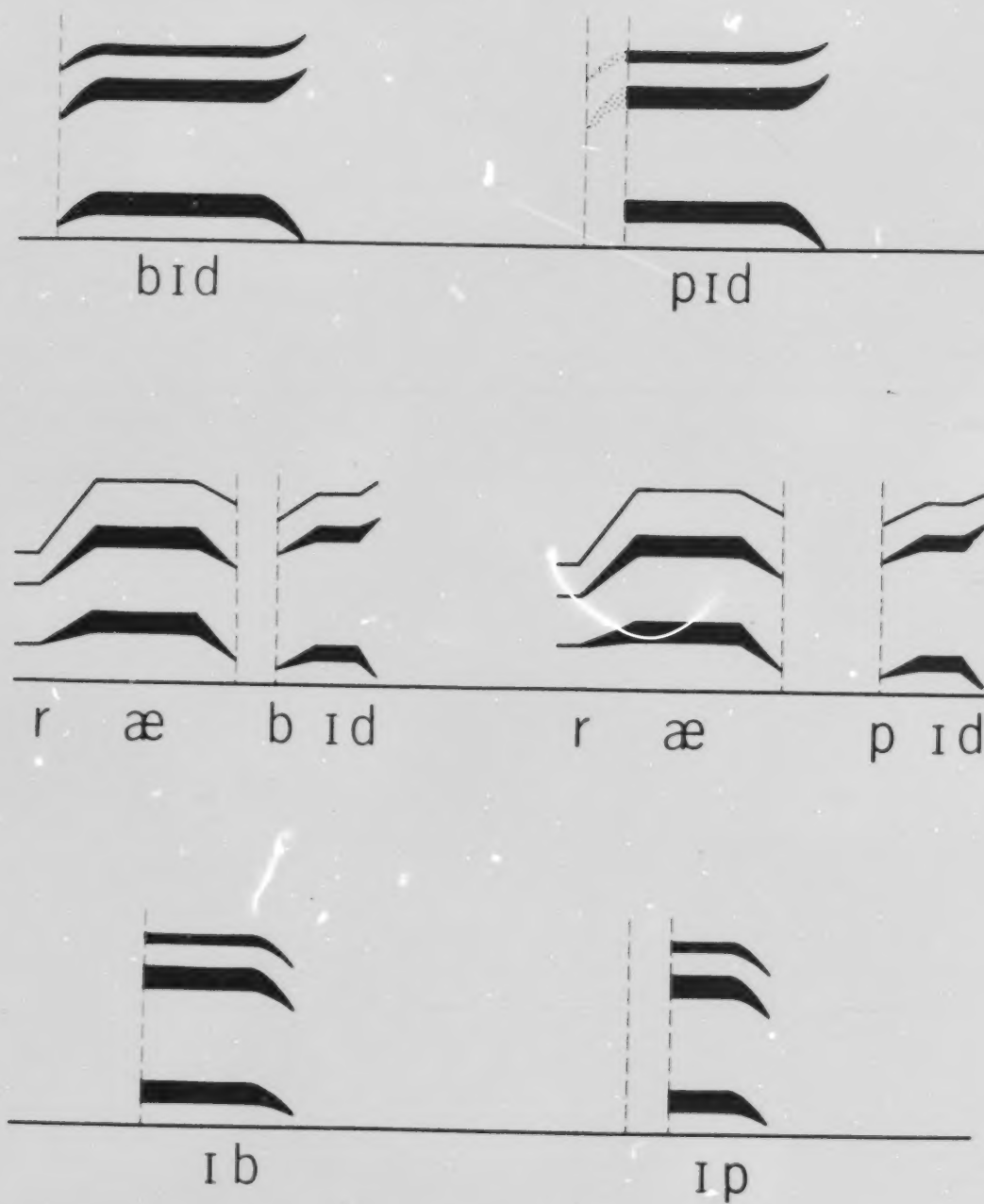


Figure 3: Schematic representation of the different acoustic cues that are appropriate for the same phonetic distinction according to its position in the syllable.

Figure 4 (Delattre, Liberman, and Cooper, 1955). We should note first that the lower (first) formant is the same for the two vowels and, indeed, for the two syllables. So we put our attention on the higher (second) formants. There we find information sufficient, in combination with the common lower formant, to tell us that the vowels are /i/ and /u/ and that the consonants are all /d/, not /b/ or /g/. We see, then, that a phonetic difference limited to the middle (vowel) segment does not produce a change in the signal that is limited to the middle portion of the sound; rather, the entire formant changes. Note especially that the transition cues for /d/ are in very different parts of the spectrum for the two syllables--high for /did/ and low for /dud/. Moreover, if the starting point of the second formant for /did/ is lowered so as to coincide with the starting point for /dud/, then listeners will most likely hear, not /did/, but /bid/. Finally, we should note that for corresponding positions in the two syllables, the transitions are opposite in direction: for /did/ they are rising in initial position and falling in final position, but for /dud/ they are falling in initial position and rising in final position. Though the acoustic cues for the /d/ have very different shapes, the underlying articulatory gesture for the consonant is much the same: a complete closure of the vocal tract produced by touching the tongue tip to the alveolar ridge. The different shapes of the acoustic cues are owing, of course, to the coarticulation of the consonant gesture with the gesture appropriate for the preceding or following vowel. As for the perception, it is as if it rationalized the context-conditioned variation in the acoustic cue and recovered the common articulation.

Very Similar Acoustic Cues Produce Very Different Phonetic Perceptions According to the Context

Having considered cases in which cues that lie on different acoustic dimensions have the same phonetic effect, we will now look at a case in which cues on a single acoustic dimension produce perceived contrasts on each of three phonetic dimensions: manner, voicing, and place. We have already seen that the presence or absence of an appropriate silence between the noise of /s/ and the syllable /lit/ will cue the phonetic distinction of manner between /split/ and /slit/. We have also noted that when those same intervals of silence are introduced between the syllables /ra/ and /bid/, they will produce the phonetic distinction of voicing between /rabid/ and /rapid/. We add now that further reductions in the duration of the silence cause the listener to hear a change of place from /rabid/ to /ratid/ (Port, 1976). It should be noted that, though the acoustic cues for all three contrasts are nominally on a single physical dimension (duration of silence), the articulatory maneuvers that produce them are not; manner, voicing, and place distinctions are different gestures made by different sets of muscles. The "place" change from /rabid/ to /ratid/ is especially interesting in that connection. As Port (1976) found, speakers normally close for a significantly shorter time (hence show a shorter silent interval) when producing the 'flap' /t/ in /ratid/ than when producing the labial /b/ in /rabid/. Given that listeners report hearing /ratid/ at the very short silent intervals (even though the spectral cues were appropriate for /rabid/), we should suppose they are once again honoring the extent to which phonetic perception is constrained by tacit knowledge of what a vocal tract can and cannot do when it makes linguistically significant gestures: it is as if the perceptu-

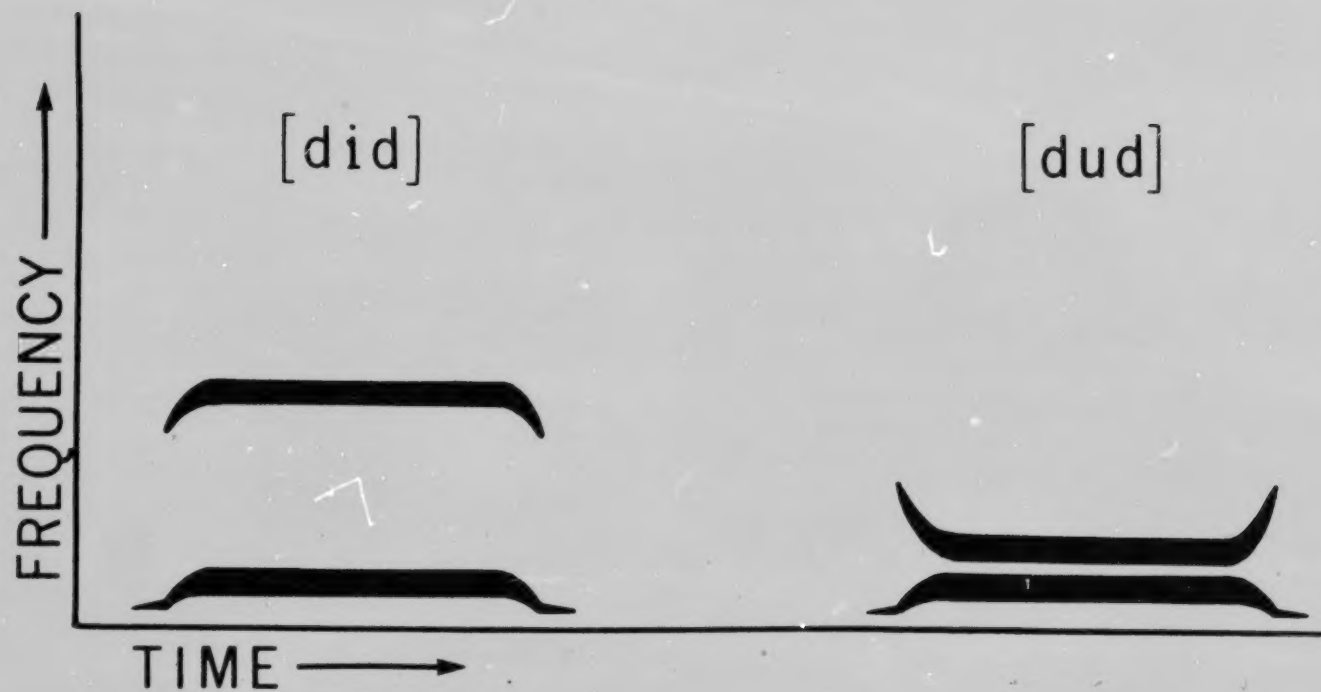


Figure 4: Schematic representation of the variation in consonant cues as a function of the vowel with which the consonant is coarticulated.

FIGURE 4

194

al system 'knew' that the speaker could not have said /rabid/ because his vocal tract had not closed long enough for that.

#### The Perception of Acoustic Cues is Different in Speech and Nonspeech Contexts

Once again we can use the example of the silent gap as a cue for the distinction between /rabid/ and /rapid/. Imagine (1) a set of (speech) stimuli in which the size of the gap is changed in relatively small steps from /rabid/ to /rapid/ and (2) a corresponding set of (nonspeech) stimuli in which the same gaps separate two bursts of noise shaped so as to correspond in amplitude and duration to the syllables of the speech stimuli. For human listeners, discrimination of the gaps is different in the two cases, being more nearly categorical in the speech stimuli (Liberman, Harris, Eimas, Lisker, and Bastian, 1961). Of course, the cue is (necessarily) in different contexts, as it is in all the speech-nonspeech comparisons that have been made. That is all the more reason to make the same comparisons with animals and infants. (See Pisoni, 1977b for further discussion).

#### Can Speech be Perceived Without Regard for its Phonetic Structure?

The various examples we have dealt with might test whether adult humans, animals, and infants do, in fact, process the speech signal so as to recover the phonetic structure that is encoded in it. But what if it is true, as some think, that the listener must recover the phones only if he wants to spell or rhyme or alliterate or do something equally elitist? When he is just listening to speech, and trying only to extract whatever meaning it may contain, does he skip the phonology altogether? Can he, in that case, deal directly with the meaningful segments--words are surely the most likely candidates--as holistic auditory patterns? Plainly, it must be possible to do that, but only within limits. We should say what two of those limits might be.

The first limit would be on the number of words that could be identified. Words do have internal phonetic structure, after all, and distinctions among them commonly depend on rearrangements of the constituent phonetic elements. But the acoustic criteria that are appropriate for one class of phonetic segments are not ordinarily appropriate for others, and, within a class, complex adjustments must be made to accommodate the variations with position, context, and speaking rate that we have referred to earlier. Thus, a procedure that works on the auditory patterns only as auditory patterns--that is, in disregard of their constituents--will presumably fail before it has identified all the patterns.

The second limit set by a failure to appreciate phonetic structure derives from the fact that in normal, fluent speech, coarticulation does not respect word boundaries. It follows then that, just as there is no acoustic criterion that divides speech into phone-size units, so also is there none that will reliably divide it into segments that correspond to words. It should be emphasized that this problem is not trivial: as mentioned earlier, a difficulty caused by coarticulation is that information about several successive phonetic segments is carried simultaneously in the acoustic signal and on the same parameter; therefore, the phonetic segments cannot be



recovered by simply cutting the continuous acoustic signal into discrete segments. Consider, then, the plight of a creature whose stored lexicon is defined only in auditory terms: applying acoustic or auditory criteria to the stream of speech, he will recover (auditory) segments that bear a random relation to the word-size segments stored in his lexicon; therefore, the number of items he must store is not equal to the several tens of thousands of words, but is rather incalculably larger than that. To get along with a store that comprises only the number of words he knows, the listener must divide the speech stream into segments whose boundaries can be coterminous with the words. Only phonetic segments--or, more properly, their underlying phonologic forms--meet that requirement. If an animal cannot recover the phonetic structure, then he should often have difficulty retrieving the words of his vocabulary from fluent speech. Hence, we should suppose that a creature may not bypass the phonetic structure if he would perceive most of what is said to him.

#### REFERENCES

- Darwin, C. J. (1967) The Perception of Speech. In Handbook of Perception: vol 7, ed. by E. C. Carterette and M. P. Friedman. (New York: Academic Press), pp. 175-226.
- Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psych. 23, 46-62.
- Delattre, P. C., A. M. Liberman, and F. S. Cooper. (1955) Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773.
- Dorman, M. F., L. J. Raphael, and A. M. Liberman. (1976) Further observations on the role of silence as a cue for stop consonants. J. Acoust. Soc. Am. 59, Suppl. 1, S40(A).
- Dorman, M. F., L. J. Raphael, A. M. Liberman, and B. Repp. Some maskinglike phenomena in speech perception. J. Acoust. Soc. Am. 57, Suppl. 1, S48(A). [Full text in Haskins Laboratories Status Report on Speech Research SR-42/43, 265-276.]
- Ganong, W. F. (1975) An experiment on "phonetic adaptation." Progress Report (Cambridge: MIT Research Laboratory of Electronics) 116, 206-210.
- Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. Lang. Sp. 1, 1-7.
- Liberman, A. M. (1974) The specialization of the language hemisphere. In The Neurosciences: Third Study Program, ed. by F. O. Schmitt and F. G. Worden. (Cambridge, Mass.: MIT Press), 43-56.
- Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
- Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial positions. Lang. Sp. 1, 153-167.
- Liberman, A. M., K. S. Harris, P. Eimas, L. Lisker, and J. Bastian. (1961) An effect of learning on speech perception: the discrimination of durations of silence with and without phonetic significance. Lang. Sp. 4, 175-195.
- Lisker, L. (1957) Closure duration and the intervocalic voiced-voiceless distinction in English. Lang. 33, 42-49.
- Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384-422.

- Pisoni, D. B. (in press a) Speech perception. In Handbook of Learning and Cognitive Processes, ed. by W. K. Estes. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Pisoni, D. B. (in press b) Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. J. Acoust. Soc. Am.
- Port, R. (1976) Influence of tempo on the closure interval cue to the voicing and place of intervocalic stops. J. Acoust. Soc. Am. 59, Suppl. 1, S41(A).
- Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of word-final consonants in American English. J. Acoust. Soc. Am. 51, 1296-1303.
- Raphael, L. J., M. F. Dorman, and A. M. Liberman. (1976) Some ecological constraints on the perception of stops and affricates. J. Acoust. Soc. Am. 59, Suppl. 1, S25(A).
- Studdert-Kennedy, M. (1976) Speech Perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press), pp. 243-293.
- Stevens, K. N. (1975) The potential role of property detectors in the perception of consonants. In Auditory Analysis and Perception of Speech, ed. by G. Fant and M. A. A. Tatham. (New York: Academic Press), pp. 303-330.
- Stevens, K. N. and A. S. House (1972) Speech perception. In Foundations of Modern Auditory Theory, vol. 2, ed. by J. Tobias. (New York: Academic Press) pp 1-62.

Further Observations on the Role of Silence in the Perception of Stop Consonants\*

M. F. Dorman,<sup>†</sup> L. J. Raphael,<sup>††</sup> and A. M. Liberman<sup>†††</sup>

ABSTRACT

Previous research has shown that silence is a necessary condition for the perception of stop consonants in such utterances as /sa/ vs. /sta/. In this case, there are cues in the vocalic portion of the syllable that are appropriate for stop-consonant manner. Hence, the silence was shown only to be a necessary, not a sufficient, cue for stop manner. In the present series of experiments we have investigated the role of silence in the perception of stops when there are no stop manner cues in the vocalic portion. The purpose was to determine the extent to which the silence cue can be not only a necessary, but also a sufficient cue.

The experiments reported here were intended to shed light on the role of silence in the perception of prevocalic stop and affricate consonants. More specifically, we were concerned with the following questions: (1) Is silence a sufficient cue to the perception of these consonants?, and (2) Is the perceptual processing of silence phonetic as well as auditory?

Previously reported experiments (Raphael, Dorman, and Liberman, 1976) have indicated that silence is a necessary but not sufficient cue to the perception of stop manner in prevocalic, as well as other, positions; that is, for example, in order to hear the stop in a contrast such as [sɛ] vs. [spɛ], we need the s-friction, a silent interval, and a vocalic segment that, in isolation, is heard as [pɛ]. Thus, silence, in this situation, may simply allow time for the [pɛ] to evade masking by the friction. But if [spɛ] could be produced, similarly, from sound segments that contain no stop--that is, if silence could be shown to be a sufficient cue--then a masking interpretation would be ruled out.

In the situations described so far, silence is seen to be a necessary but not sufficient cue for the perception of a stop manner; that is, in order

---

\*This is a slightly revised version of a paper presented at the 91st meeting of the Acoustical Society of America, Washington, D.C., 4-9 April 1976.

<sup>†</sup>Also Arizona State University, Tempe, Arizona.

<sup>††</sup>Also Herbert H. Lehman College and the Graduate School of the City University of New York.

<sup>†††</sup>Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

to hear the stop in the [se-spe] contrast, for example, we need the s-friction, a silent interval, and a vocalic segment that, in isolation, is heard as [pe]. Thus, silence, in this situation, may simply allow time for the [pe] to evade masking by the friction. But if [spe] could be produced, similarly, from sound segments that contain no stop--that is, if silence could be shown to be a sufficient cue--then a masking interpretation would be ruled out.

It is of interest, then, to ask whether there are situations in which silence is, in fact, a sufficient cue for perception of stop manner. We have long suspected that there are, given the finding many years ago that [slit] can be changed to [split] by inserting an appropriate amount of silence between the [s] friction and the onset of the vocalic segment (Liberman, Harris, Eimas, Lisker, and Bastian, 1961). That experiment was done by cutting apart and reassembling the frictional and vocalic portions of the syllable [slit] and recorded on magnetic tape; hence there may have been transients at the onset of the vocalic portion that were more or less sufficient, given the necessary amount of silence preceding them, to produce perception of the stop. We have, therefore, repeated and extended that experiment, but in a way more suitable to our purposes. To avoid having stop-relevant cues in either the frictional or vocalic components, we recorded, separately, the hissing noise of [s] and the syllable [lit]. Having determined then that there was no perceptible [p] in the noise or in the syllable [lit], we used the pulse-code-modulation (PCM) system at Haskins Laboratories to put the noise before the [lit] with various intervals of silence in between. We then randomized the resulting patterns, and presented them to listeners for judgment. The results are shown in Figure 1. We see that at silent intervals less than 70 msec, listeners reported [slit]--that is, they did not hear a stop--but when the silent interval was greater than 70 msec, they reported [split]--that is, they heard the stop consonant. We may conclude, then, that in this case silence is a sufficient cue for perception of the stop manner.

How do we interpret the role of silence in the experiment just described? It seems to us plain that the silence provides information, not time to evade masking. The information is that the speaker either did or did not close his vocal tract appropriately for production of the [p] segment in the syllable [split]. But, to use that information in order to arrive at the perception of a stop would appear to require a process that is not merely auditory but is, rather, more abstractly phonetic.

If the role of silence is, in fact, to provide information about vocal-tract closure necessary for the production of a stop, we might suppose that a stop would not be heard at silent intervals greater than those produced by normal articulation. To test that expectation, we carried out a second experiment in which we extended the silent intervals of the first experiment all the way out to 650 msec. The outcome is shown in Figure 2. We see there that, as in the first experiment, [slit] is heard at relatively short intervals of silence, and [split] at somewhat longer intervals. We also see that at still longer intervals of silence the listeners once again hear [slit]. Thus, intervals of silence much longer than those that characterize the stop closure do not produce the perception of the stop.

Let us turn now to another phonetic contrast for which the presence or absence of the stop closure is again a distinguishing articulatory feature,



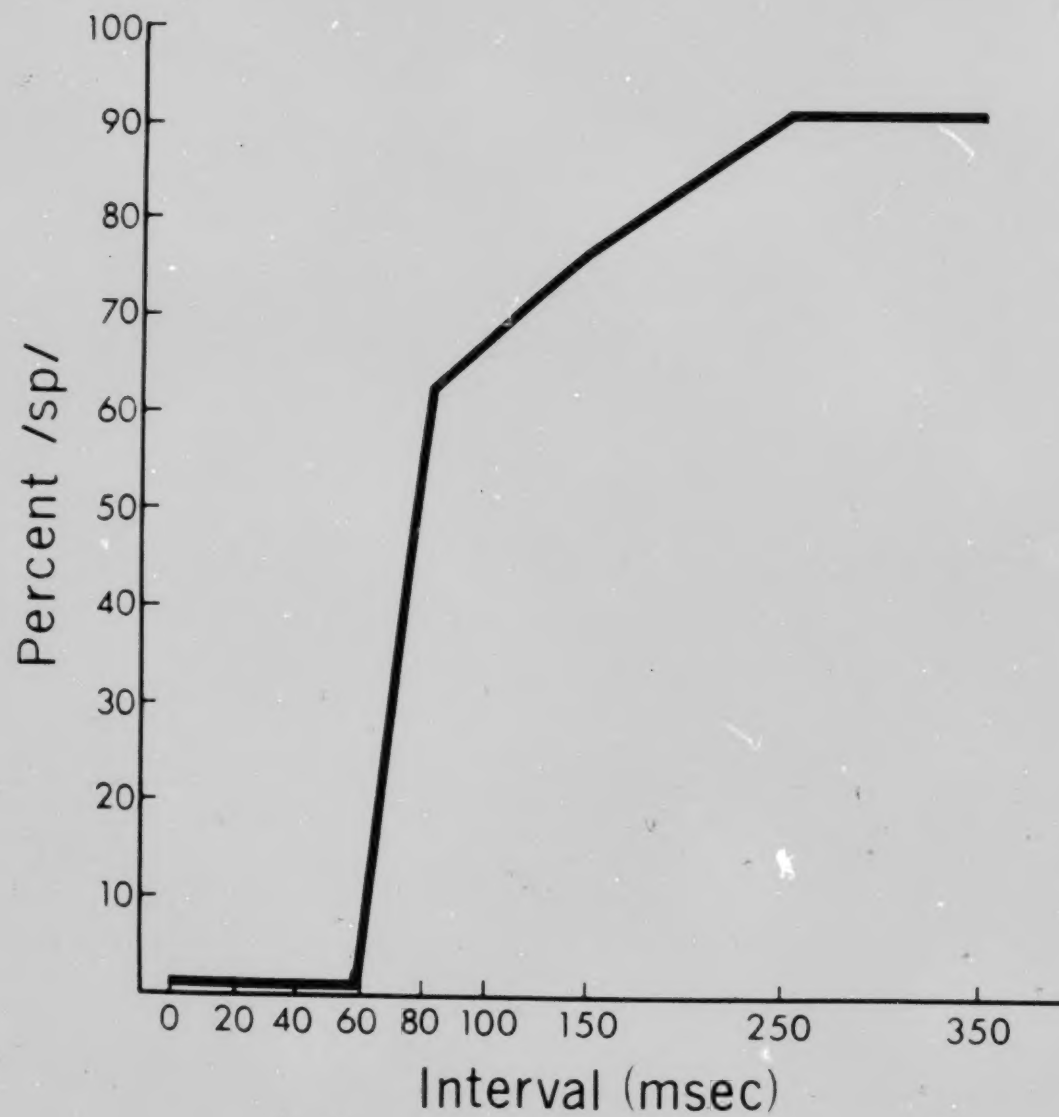


Figure 1: Percent identification of "split" as a function of the interval between /s/ and /lit/.

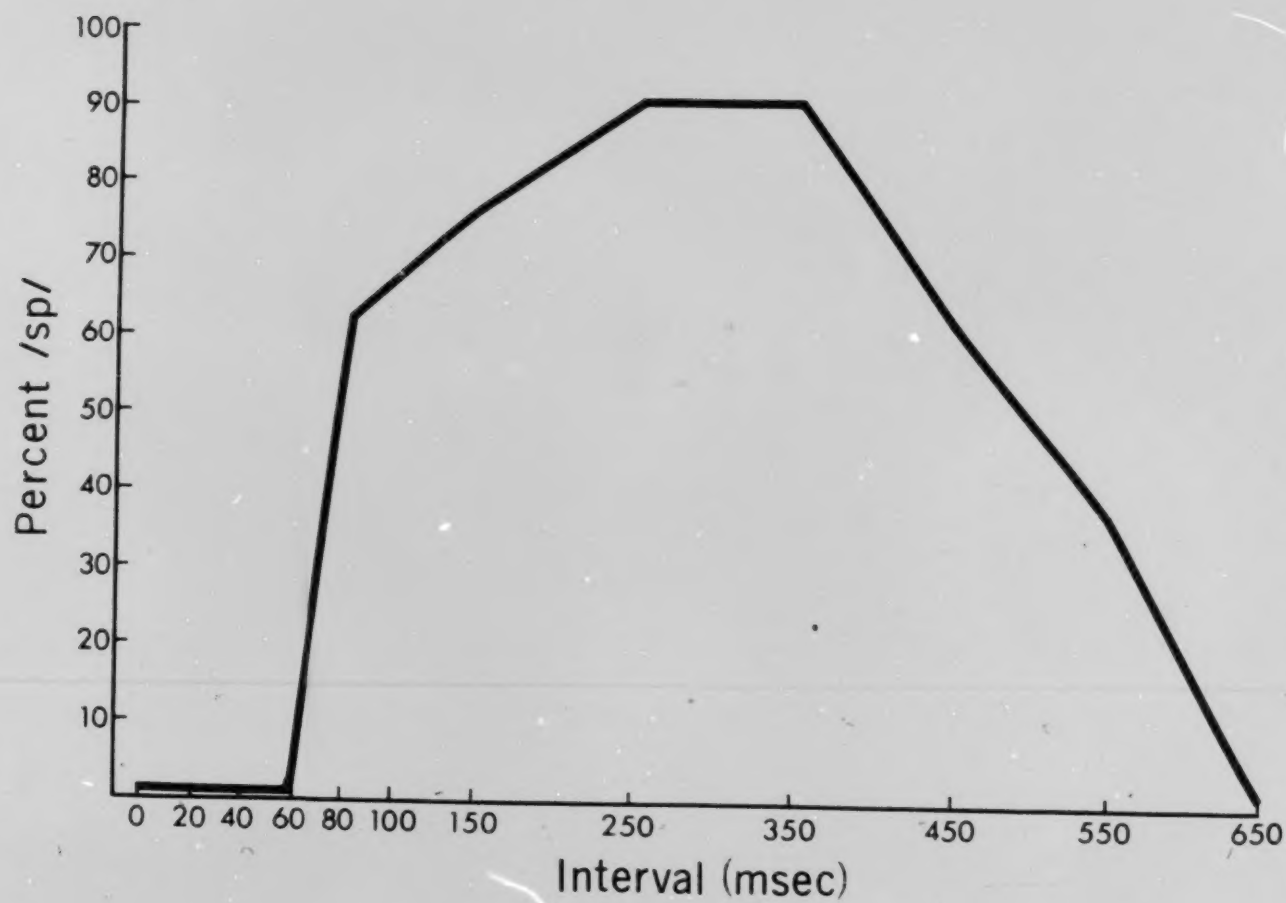


Figure 2: Percent identification of "split" as a function of the interval between /s/ and /lɪt/.

but which differs from [slit] vs. [split] in two respects: the stop closure occurs between syllables, not within a syllable, and the contrast is between fricative and affricative. The example is the difference between the utterances "Please say shop" and "Please say chop." Is silence a sufficient cue here, too? To find out, we performed the following experiment.

Having recorded the utterance "Please say shop," we separated "please say" from "shop" and then recombined them with silent intervals between "say" and "shop" that varied from 0 to 150 msec. The resulting patterns were randomized and presented to listeners with instructions to identify each one as "please say shop" or "please say chop." The results are shown in Figure 3. We see that at intervals of less than about 50 msec, listeners reported hearing "please say shop," while at longer intervals they heard "please say chop." Thus, silence is a sufficient condition for the perception of the affricate in absolute initial position in the syllable.

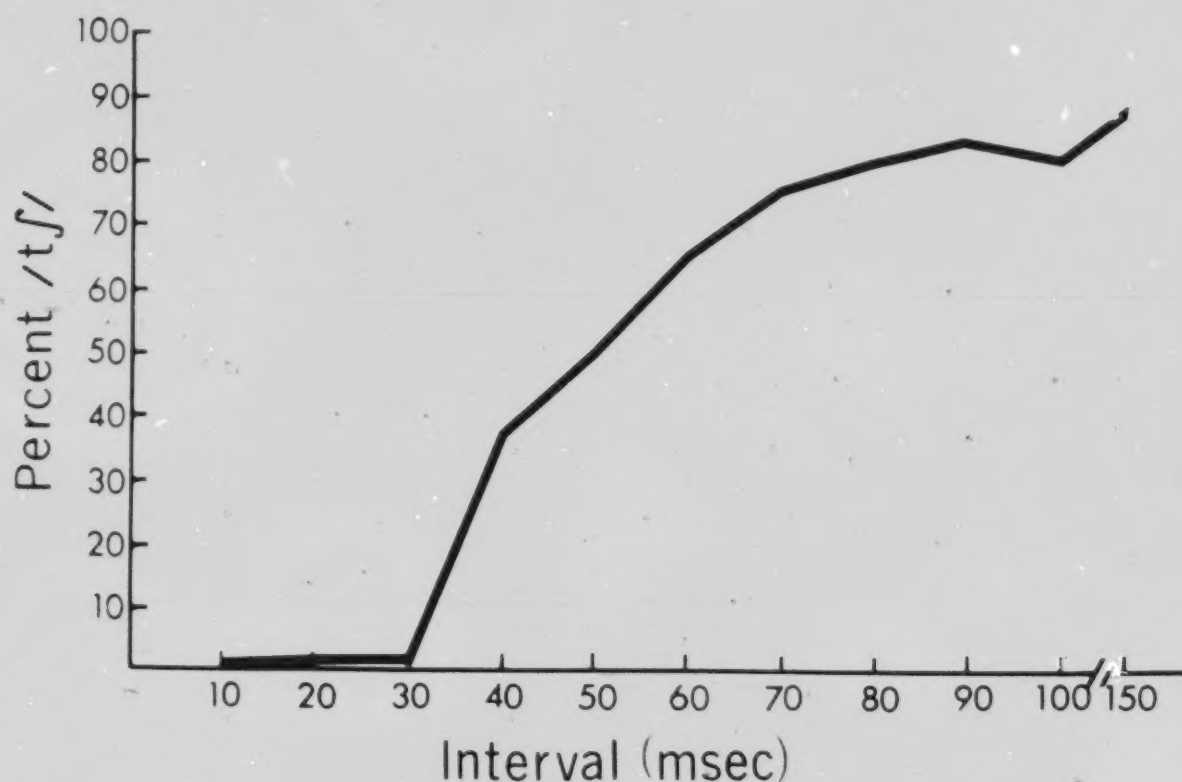


Figure 3: Percent identification of "chop" as a function of the interval between the carrier phrase "please say" and "shop."

We suppose that here, too, silence provides the (phonetic) information that the speaker either did or did not close his vocal tract in the manner necessary for the production of the affricate. And here, too, we thought it

of interest to determine whether the segment would not be perceived if the silent interval were longer than that produced by normal articulation. To find out, we extended the time intervals between "please say" and "shop" and found, as shown in Figure 4, that at very long silent intervals the listeners do indeed once again hear the fricative (rather than the affricate) in "please say shop."

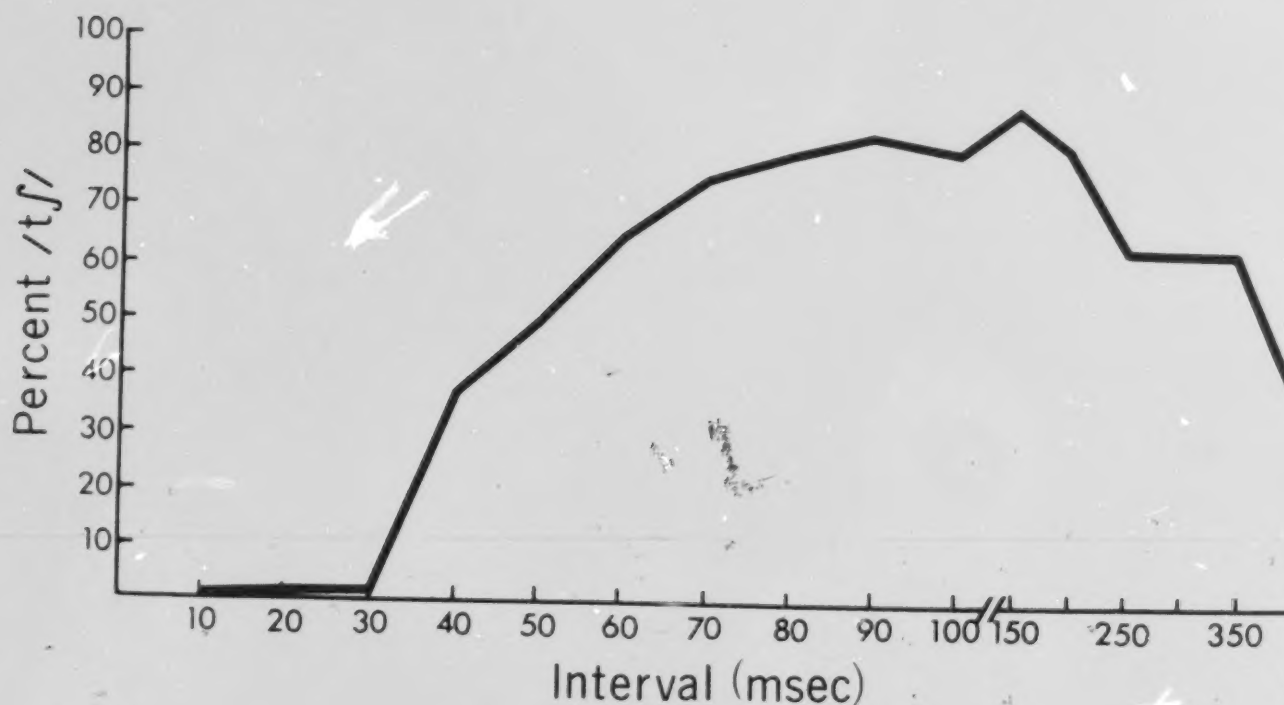


Figure 4: Percent identification of "chop" as a function of the interval between the carrier phrase "please say" and "shop."

The last experiment we shall report here was designed to determine whether the effect of the silent-interval cue varies according to the tempo at which the precursive phrase "please say" is uttered. In one sense, this experiment is premature, at least from our point of view, since we do not now know how the duration of the speaker's closure is affected by variations in the rate at which he speaks. However, we do know from perceptual experiments that the effect of certain other temporal cues does vary with speaking tempo, and, since such findings are relevant to questions about how the cues are processed, we plunged ahead.

To determine the effect of tempo, we recorded "please say" at two rates, one slow, the other fast. We then prefixed the resulting phrases to "shop"



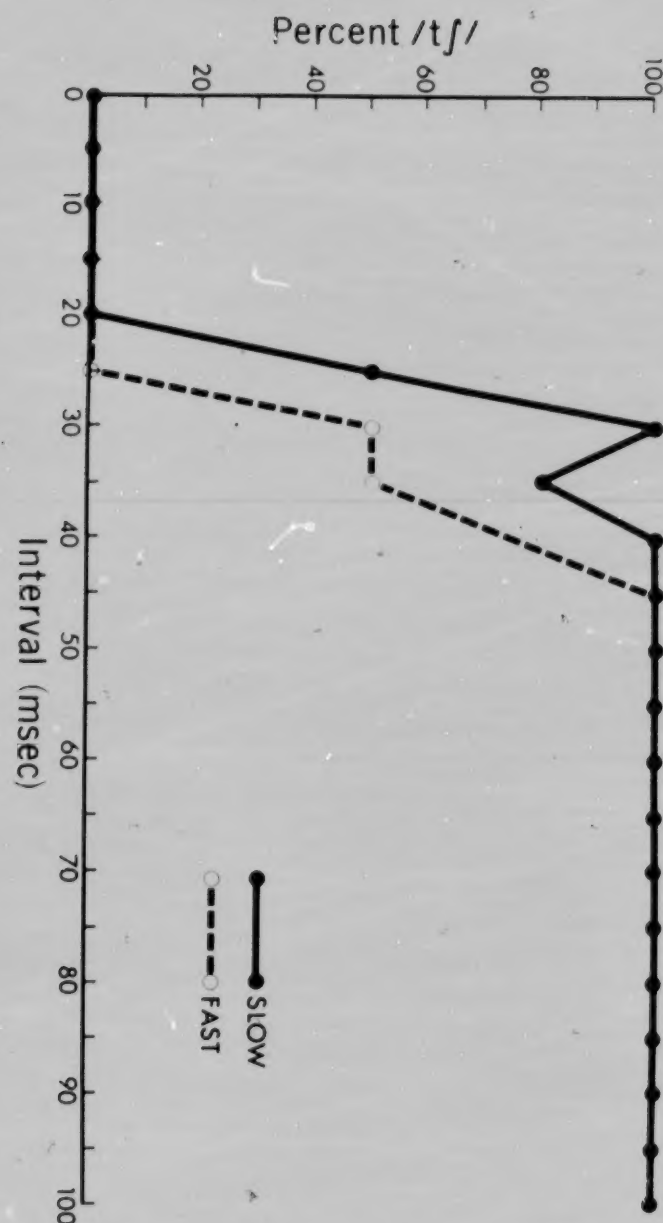


Figure 5: Percent identification of "chop" as a function of the interval between the carrier phrase "please say" and "shop" for carrier phrases uttered at a slow and fast rate.

at silent intervals of 0 to 100 msec between the end of "say" and the beginning of "shop." These patterns were randomized and presented to listeners for judgment as "please say shop" or "please say chop."

The results are shown in Figure 5. The solid curve is the labeling function for the slow condition. We find the boundary at about 23 msec. The boundary for the fast condition, represented by the dashed line, lies at about 30 msec. Thus, there is an effect of rate, but its direction indicates that at the faster speech rate our listeners needed a longer silent interval before they perceived the affricate in "chop." At the present time we cannot interpret that finding with confidence, since, as we said, we have not measured the effects of speaking rate on articulation of the various parts of the utterance. We can only remind ourselves that the duration of the friction is also a cue for the distinction between fricative and affricative--longer for fricative, shorter for affricative--and then suppose that, given the fast precursive phrase, the duration of the friction would seem longer and hence would bias the perception toward the fricative. Conceivably, that overrode the contrary bias that might have been expected from the silence cue. In any case, we may conclude that the effect of the silence cue is, in fact, sensitive to speaking rate. To interpret that result, we shall have to invoke processes that go beyond the psychophysics of gap perception.

We should summarize. We have found that silence can be a sufficient cue for the perception of the stop (in [slit] vs. [split]) and the affricate (in [shop] vs. [chop]): Moreover, the effective cue is not just any duration of silence, nor is it precisely that duration which is perceptibly different from no silence; it is, rather, a range of durations that appear to match reasonably well the silences that result from the vocal tract closures when the stop and affricate are produced. In the case of the affricate-fricative distinction, the effect of the silence cue was also found to be sensitive to the rate of articulation. We suppose that the role of silence is to provide information about the presence or absence of the vocal-tract closure that marks the production of the stop and affricate phones. Accordingly, we assume that the perceptual processing of the silence cue is not only auditory but also phonetic.

#### REFERENCES

- Ainsworth, W. A. (1973) Durational cues in the perception of certain consonants. In Proceedings of the British Acoustical Society 2, 1-4.
- Gerstman, L. (1957) Perceptual dimensions for the friction portions of certain speech sounds. Ph.D. thesis, New York University.
- Liberman, A. M., K. S. Harris, P. Eimas, L. Lisker, and J. Bastian. (1961) An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. Lang. Speech 4, 175-195.
- Port, R. (1976) Influence of tempo on the closure interval cue to the voicing and place of intervocalic stops. JASA 59, S41(A).
- Raphael, L. J., M. F. Dorman, and A. M. Liberman. (1976) Some ecological constraints on the perception of stops and affricates. J. Acoust. Soc. Am. 59, Suppl. 1: S25(A).
- Summerfield, A. Q. (1975a) Information-processing analyses of perceptual adjustments to source and context variables in speech. Unpublished Ph.D. thesis, Queen's University, Belfast.
- Summerfield, A. Q. (1975b) How a full account of segmental perception

depends on prosody and vice versa. In Structure and Process in Speech Perception, ed. by A. Cohen and S. G. Nooteboom. (New York: Springer-Verlag), pp. 51-67.

# Perception of Implosive Transitions in VCV Utterances

Bruno H. Repp

## ABSTRACT

The present experiments demonstrate that, in making judgments (same-different or classification) about the medial stop consonants in synthetic vowel consonant vowel (VCV) utterances, listeners can utilize the information provided by the implosive transitions of the stop, even though implosive transitions are not heard as a separate phonemic event and despite their relatively weak cue value. This conclusion is based on the finding that reaction times (RTs) to the medial stop increase little or not at all as the closure period of the stop is lengthened. The task of attending to the vowel consonant (VC) portions of VCV utterances was apparently facilitated by intermixing VCV and VC stimuli and by randomizing the closure duration variable. Nevertheless, few subjects were able to completely ignore the information following the closure period (explosive transitions and final vowel). A paradoxical effect obtained in one study (an influence of the final vowel on RTs supposedly based on the VC portion of the stimuli, even at long closure durations) needs to be clarified in further research. Ancillary studies determined the closure period necessary to hear double (geminate) voiced stop consonants to be in the vicinity of 200 msec, and showed that, as closure duration is shortened, the discrimination of VCV stimuli with and without implosive transitions becomes exceedingly difficult and approaches chance at a closure duration of 65 msec.

## INTRODUCTION

Repp (1975, 1976) reported two studies of "coperception" in VCV utterances. Coperception was defined as the influence of one segment on the perception of another segment in an utterance, the segments being defined at the phonemic level. The dependent variable was the reaction time (RT) of same-different judgments about the medial stop consonants in two successive VCV utterances. The final vowel was variable and constituted the irrelevant

---

Acknowledgement: Experiment I was conducted while I was a research associate at the University of Connecticut Health Center and supported by NIH Grant T22 DE00202 to that institution. It was made possible by the extraordinary hospitality of Haskins Laboratories and of its president, Alvin Liberman. I thank him, Quentin Summerfield, and Alice Healy for their valuable suggestions and their interest in my work.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]



context whose influence on the same-different latencies was investigated.

The medial stops consisted of (1) implosive transitions from the initial vowel that were independent of the final vowel, (2) a silent closure period, and (3) explosive transitions into the final vowel. A typical stimulus is illustrated in Figure 1; it may be considered as a VC portion followed by a CV portion after a silent interval. Given that the implosive transitions alone are sufficient to identify the stop consonant, the question was whether the stop would be "co-perceived" with the final vowel via the explosive transitions. Pisoni and Tash (1974) and Wood and Day (1975) had earlier shown that in CV syllables where the stop is cued only by the vowel-dependent explosive transitions, the consonant is co-perceived with the vowel.

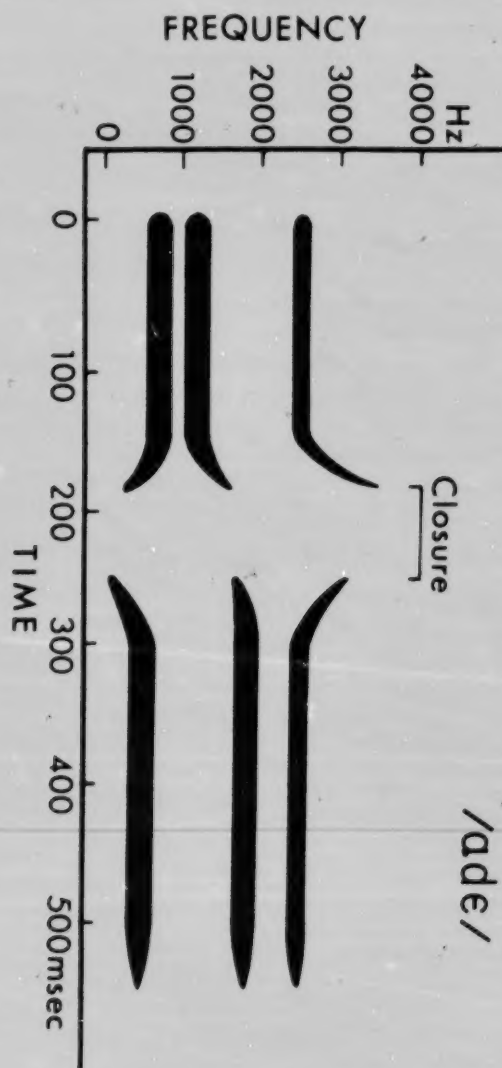


Figure 1: Schematic spectrogram of a typical VCV stimulus with a 65-msec closure period.

My first study (Repp, 1975) showed that the same result is obtained with VCV utterances: latencies of "same" judgments about the medial stops in two successive VCVs were faster when the final vowels were the same than when they were different. (A tendency in the opposite direction for "different" latencies was not as striking and constituted only secondary evidence.) This result suggests that the subjects were not able to take advantage of the implosive transitions (that did not vary with the final vowel), but instead relied exclusively on the explosive transitions (that did vary with the final vowel), as in CV syllables.

This result is in accord with experiments that have shown that the explosive transitions determine the perception of the medial stop in VCVs when implosive and explosive transitions are artificially brought into conflict and the closure interval is sufficiently short (Abbs, 1971; Dorman, Raphael, Liberman, and Repp, 1975; Fujimura, 1975). Dorman et al. varied the duration of the closure period following the implosive transitions and found that 50-80 msec are necessary for conflicting implosive transitions to be perceived as a separate phoneme, that is, for an utterance to be perceived as VC<sub>1</sub>C<sub>2</sub>V rather than as VC<sub>2</sub>V. A much longer interval (not precisely determined by Dorman et al.) was needed to perceive as a separate event implosive transitions that cued the same place of articulation as the explosive transitions, that is, for the utterance to be heard as VC<sub>1</sub>-C<sub>1</sub>V, where C<sub>1</sub>-C<sub>1</sub> is a geminate stop. Since the closure interval in my stimuli was only 50 msec long, the implosive transitions were obviously not perceived as a separate event; instead, they were perceptually integrated with the following explosive transitions. Massaro's (1975) account of Abbs' (1971) results ascribes the phenomenon to backward recognition masking: at short closure periods, there is not enough processing time available for the implosive transitions to be perceived as a separate unit. This view is also in agreement with my finding.

In a second study (Repp, 1976), I again used VCV utterances in a same-different paradigm, but I varied the duration of the closure period from 65 to 215 msec. I predicted that coperception of the stop with the final vowel would disappear at longer closure period durations. Specifically, I expected that the effect would vanish as soon as geminate consonants are perceived, which I expected to happen at the longest closure period duration (215 msec). The results did not confirm my expectations; they indicated that the subjects continued to rely only on the explosive transitions, even at the longest closure duration, despite instructions to respond as fast as possible and not to wait for the end of the second utterance in a pair before responding.

This experiment omitted two important control conditions. First, it was not tested how well the implosive transitions could be discriminated in isolation (that is, when only the VC portion of the stimuli is presented). The result of the study would be trivial if the subjects had not been able to discriminate the relevant portion of the stimuli. Although this seemed unlikely (and several subjects were tested later and showed that they could discriminate /ab/ from /ad/), the possibility remained that very slow latencies were associated with these discriminations, so that the listeners found it more advantageous to rely on the perhaps more discriminable CV

portions of the VCV utterances. The other important omission was that I did not test whether the subjects actually heard geminates at the longest closure period. Instead, I relied mainly on my own impressions, since at that time I was not aware of any relevant data in the literature. If the subjects actually did not hear geminates, and if my hypothesis that coperception and geminate perception are mutually exclusive was correct, the negative result of the experiment is no longer surprising. An additional factor may have been the blocked presentation of the different closure durations. The present two experiments rectify these shortcomings, and in addition, provide new data on the perception of implosive transitions in VCV utterances.

### EXPERIMENT I

This experiment contained three tasks. Task 1 simply consisted of classifying the VC portions /ab/, /ad/ presented in isolation. Both speed and accuracy were measured. Task 2 likewise required a choice between /ab/ and /ad/, but here these VC portions were presented in VCV context, that is, they were followed by CV portions after a variable closure period. A few VC tokens were also included. Closure duration was randomized and varied over a wide range (65 to 365 msec) that was certain to include stimuli perceived as containing geminates. (The single-geminate boundary was formally determined in Experiment II, and found to be slightly above 200 msec.) With reaction time (RT) being measured from the beginning of the closure period, latencies should be independent of closure period duration if the subjects are able to "pick out" and respond to the implosive transitions, but they should increase linearly with closure duration if the subjects rely only on the explosive transitions. The most likely outcome was considered to be intermediate between these two extremes. At short closure periods, the listeners might rely on the explosive transitions, so that the latencies would increase with closure duration, but the latency function would level off as the single-geminate boundary is approached: from there on the subjects would rely on the implosive transitions and the latencies would be independent of closure duration.

Task 3 was an extension of the previous same-different RT experiment. In addition to pairs of VCV stimuli with various closure durations, VC-VCV, VCV-VC, and VC-VC pairs were also included in a completely randomized design. It was expected that the inclusion of VC stimuli would facilitate or even force the listeners' attention to the VC portion of the VCV stimuli. Failures to follow the instructions and reliance on the explosive transitions in VCVs should lead to difficulties ("surprise" reactions and slower latencies) with VCs. The principal questions concerned the absolute RTs and the coperception effect of the final vowel on "same" RTs as a function of closure duration.

### METHOD

#### Subjects

Nine paid volunteers participated. Some of them had participated in earlier experiments with synthetic speech, but all were relatively inexperi-

enced in RT tasks. In addition, three experienced listeners (the author and two colleagues at Haskins Labs) took the experiment, with the exception that less data were collected in Task 3 than for the naive subjects.

### Stimuli

The stimuli were the VCV utterances /abɛ/, /abi/, /adɛ/ and /adi/, synthesized on the Haskins Laboratories parallel formant synthesizer, plus the two VC syllables /ab/ and /ad/ that represented the initial portions of the VCVs. The VCVs consisted of a 185-msec VC portion and a 300-msec CV portion, separated by variable silent closure interval (cf. Figure 1). The closure durations used were 65, 165, 265, and 365 msec.<sup>1</sup>

There were three experimental tapes corresponding to the three tasks. Tape 1 contained 50 randomized VC syllables, with onset-to-onset intervals of 3.85 sec. Tape 2 contained a sequence of 90 VCV utterances made up of five different successive randomizations of all 18 stimuli used in the experiment. The onset-to-onset interval was again 3.85 sec. Tape 3 contained two blocks of 132 stimulus pairs each. The 132 pairs comprised all 64 VCV-VCV combinations with identical closure durations (VCVs with different closure durations were never paired with each other), all 32 VC-VCV combinations, all 32 VCV-VC combinations, and the 4 VC-VC combinations. Their sequence was completely random. The stimulus onset asynchrony within a pair was constant at 1 sec. The interval between the onset of the second stimulus in a pair and the onset of the first stimulus of the next pair was constant at 3.85 sec.

### Procedure

Each subject listened first to Tape 1 with instructions to press one response key for /ab/ and the other key for /ad/ (Task 1). The response-hand assignment was counterbalanced across subjects. This was followed by Tape 2 for which the task remained the same, except that it was emphasized to respond as quickly as possible and to ignore whatever followed the initial VC part of each utterance (Task 2). The remainder of the experiment was taken up by the same-different task (Task 3). Again, the subjects were instructed to focus on the initial VC portions which alone were relevant to the decisions to be made and to ignore the final vowel. Each subject listened to Tape 3 twice, that is, to four blocks of 132 pairs (ca. 12 minutes per block). They responded "same" with their preferred hand (the right hand for all but one subject). All tasks were preceded by examples selected randomly

---

<sup>1</sup>These stimuli are the same as in Repp (1976), except for the longer closure durations. In the earlier report, the duration of the VC portion was mistakenly given as 15 msec longer and the durations of the closure periods as 15 msec shorter than was actually the case. Also, these values were off by another 5 msec in all stimuli beginning with /ad/. This latter difference was unintended but certainly inconsequential; it was eliminated in Experiment II.



from the tape. All stimuli were presented binaurally at a comfortable listening level. All other details of procedure were the same as in Repp (1976).

### Analysis

All RTs were measured from the offset of the implosive transitions (that is, the onset of the closure period). The analysis was conducted on mean latencies, omitting errors, latencies shorter than 165 msec, and obvious outliers. For the statistical analysis of Task 3, Bock's (1975) multivariate method of treating repeated-measurements data that yields a multivariate estimate of  $F$  with very conservative degrees of freedom (denoted here by  $F$ ), was used. Other  $F$  values reported here were obtained by standard ANOVA procedures. The data of the three experienced subjects were kept separate from those of the naive subjects.

## RESULTS

### Task 1

The two VC syllables were easily discriminated by all subjects. The overall error rate was 1.6 percent; 2.2 percent (5 errors) for /ab/ and 0.9 percent (2 errors) for /ad/. An additional 6 responses were excluded because of unusually long latencies. RTs were faster to /ad/ (405 msec) than to /ab/ (482 msec). This difference was shown by eight of the nine subjects and was significant ( $F_{1,8} = 9.59$ ,  $p < .02$ ). A difference in the same direction was shown by two of the three experienced subjects whose average latencies were 247 msec for /ab/ and 242 msec for /ad/. The faster RTs for /ad/ were probably due to a difference in the quality of the synthetic stimuli: While /ab/ had a "quiet" offset, /ad/ had a characteristic "ringing" sound at offset (caused probably by the high frequency of the third formant) that permitted it to be identified more easily.

### Task 2

Figure 2 shows the results for all individual naive subjects (numbered 1-9) and the three experienced subjects (BHR, GMK, and AQS). Average choice RTs are shown as a function of closure duration. The extreme right data points represent the latencies for VC stimuli (VCVs with an infinitely long closure period, as it were); they are based on only 10 observations for each subject, while all other data points are based on 20 observations (minus errors and outliers).

It can be seen that RTs increased with closure duration for most naive subjects. The overall increase was significant ( $F_{3,24} = 8.05$ ,  $p < .01$ ). Individual functions varied a great deal: some subjects showed monotonic increases (subjects 2, 4, 6, 7, and perhaps 8), while the remaining subjects showed nonmonotonic functions. One subject (3) showed little variation, and two additional subjects (1 and 5) would have shown a flat function if they had not produced abnormally long latencies at the 265-msec closure duration. Only a single subject (4) showed a latency function with a slope close to

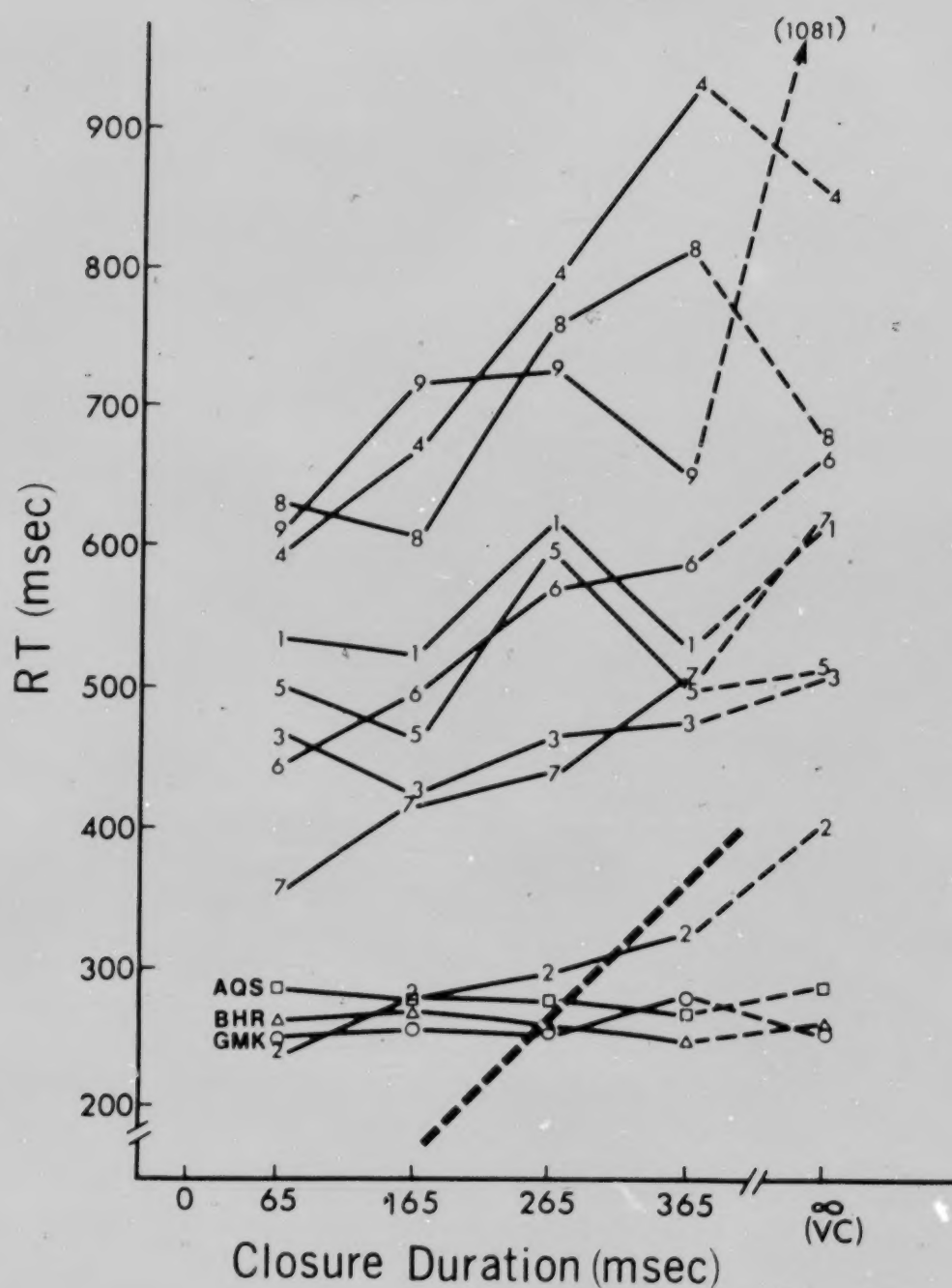


Figure 2: Average choice-reaction times (RTs) as a function of closure duration. Data for individual inexperienced (1-9) and experienced (BHR, GMK, AQS) subjects. The heavily dashed line at the lower right indicates the end of the closure period on the reaction time scale (ordinate).

unity. (A heavily dashed line with this slope is drawn in the lower right-hand part of Figure 2).

For all subjects, VC latencies tended to be slower than the average VCV latencies, and for most subjects, slower than the VCV latencies at the longest closure period. However, because of large individual variations, the latter difference did not reach significance. All subjects gave much slower responses to VC syllables in Task 2, where they were embedded in a large majority of VCV stimuli, than in Task 1, where only VC stimuli were presented (661 vs. 444 msec;  $F_{1,8} = 18.00$ ,  $p < .01$ ).

The three experienced subjects, on the other hand, had completely flat latency functions--RT was not influenced by closure duration. Also, VC latencies were of the same magnitude as VCV latencies. The task effect on VC latencies was present, however, for all three listeners: the average RTs to VCs were 270 msec in Task 2, but 245 msec in Task 1. The heavily dashed line in the lower right-hand corner of Figure 2 represents the end of the closure period on the reaction time scale. All data points lying below this line represent responses that, on the average, were made before the CV portion of the stimulus had even begun. Thus, these responses must be completely independent of the duration of the closure period.

The difference between /ab/ and /ad/ latencies observed in Task 1 was no longer present in Task 2. Instead, there was a curious interaction of the final vowel with the closure duration effect ( $F_{3,24} = 8.01$ ,  $p < .01$ ): at the two shortest closure periods, RTs were faster when the final vowel was /ε/; at 265 msec there was a large difference in the opposite direction, and at the longest closure period there was no difference at all.

The average error: percentage in Task 2 was 5.9, with individual rates ranging from 2.2 percent to 13.3 percent. (An additional 1.6 percent of the responses were excluded as outliers of the latency distributions.) Errors decreased as closure duration increased; the respective percentages were 10.0, 6.1, 5.0, and 2.8, and 5.6 percent for VC syllables. Apart from one subject who made almost all her errors with /ade/, there was no obvious relationship to the particular stimuli.

### Task 3

The results are shown in Figure 3. RTs are shown in the upper panels and error rates in the lower panels. Let us first consider the results for VCV-VCV pairs only (left-hand panels). "Same" (S) (RTs) and errors (note that "S" errors represent incorrect "different" responses) are coded as circles connected by solid lines, "different" (D) RTs and errors as triangles connected by broken lines. The vowel context is coded by filled (identical vowels, =) vs. open (different vowels, #) symbols, as indicated by the subscripts to the S and D labels.

First of all, it can be seen that RTs were fairly independent of the closure interval ( $F_{3,6} = 2.10$ ,  $p < .10$ ). Only between the two longest intervals, "same" latencies seemed to increase while "different" latencies

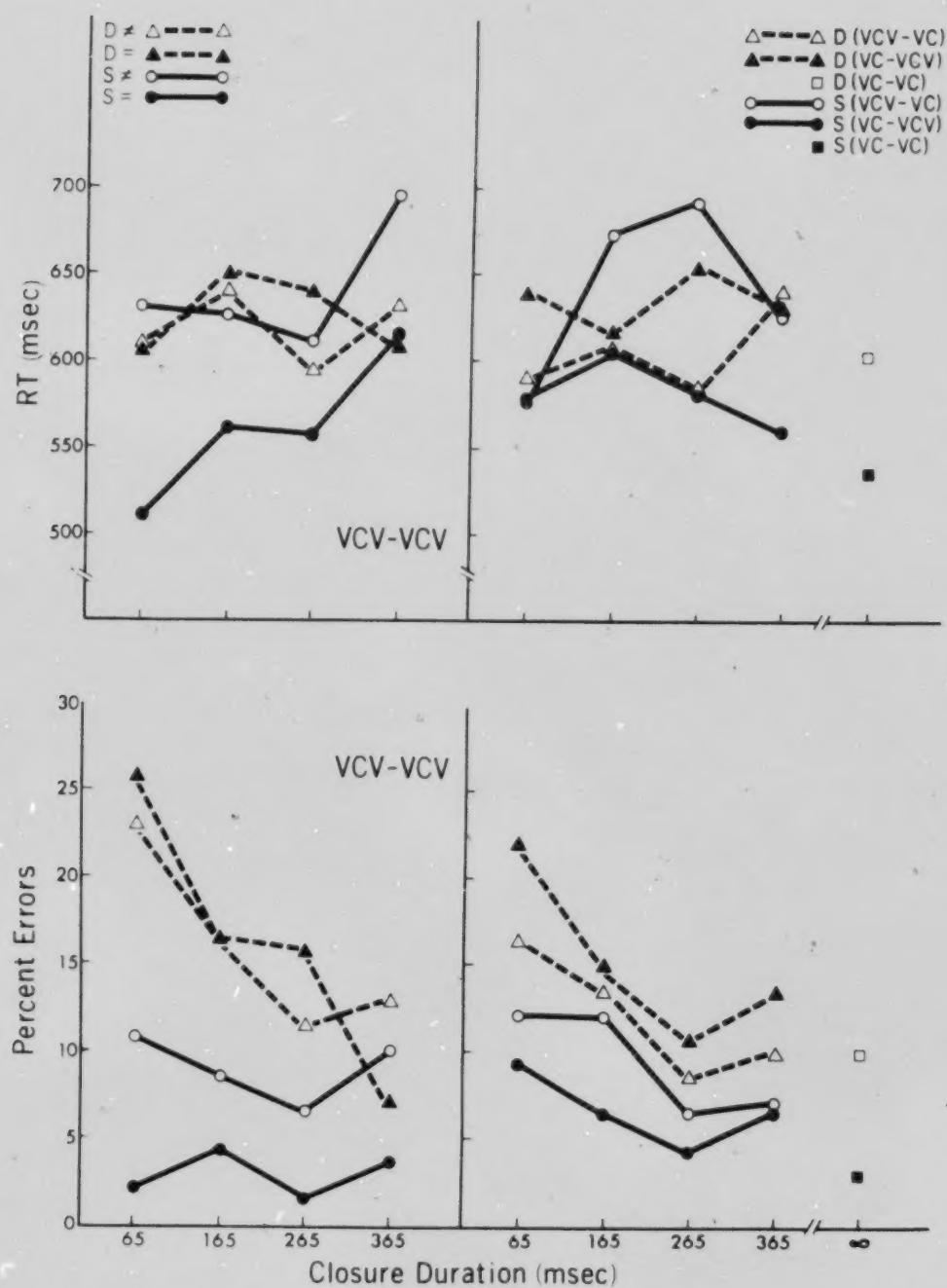


Figure 3: Left-hand panels: Average reaction times (top) and error percentages (bottom) for "same" (S) and "different" (D) judgments in identical (=) and different (≠) final-vowel contexts as a function of closure duration. Right-hand panels: Average reaction times (RTs) (top) and error percentages (bottom) for "same" (S) and "different" (D) judgments in VC-VCV and VCV-VC pairs as a function of the closure duration of the VCV stimuli. Results for VC-VC pairs are shown at the very right.

217



remained constant; the relevant component of the closure duration by same-different interaction was significant ( $F_{1,8} = 8.51, p < .02$ ), but not the overall interaction. Second, it is evident that there was a clear effect of vowel context on "same" latencies, with faster latencies in identical contexts than in different contexts. This effect was present at all closure durations. "Different" latencies did not show any systematic context effect. This was reflected in a significant same-different by vowel context interaction ( $F_{1,8} = 14.9, p < .0006$ ). The context effect seemed to decrease somewhat with practice (blocks 1 and 2 versus blocks 3 and 4:  $F_{1,8} = 6.66, p < .04$ ), together with the RTs ( $F_{1,8} = 6.11, p < .04$ ). There were no other interactions with practice. The same-different main effect did not reach significance. There were not sufficient data to investigate effects tied to individual stimuli.

The overall error rate was 10.9 percent (this includes some correct but unusually long latencies). Individual error rates ranged from 0.8 percent to 25.8 percent. The error pattern (lower left-hand panel in Figure 3) showed similarities with the latency pattern: again, the copercception (or vowel context) effect was observed for "same" trials (that is, incorrect "different" responses), with no systematic effect on "different" trials. Errors on "same" trials were clearly less frequent than on "different" trials. In addition, the former were not related to closure duration, while the latter decreased quite rapidly as the closure periods got longer. In other words, the errors decreased where the latencies were constant, and they were constant where the latencies increased.

The right-hand panels of Figure 3 illustrate the results for stimulus pairs that included at least one VC syllable. Again, circles and solid lines represent "same" latencies and errors, triangles, and dashed lines, "different" latencies and errors. The other distinction here is the temporal order of the two stimuli in a pair, that is, whether the VC syllable was first (filled symbols) or second (open symbols). In addition, latencies to pure VC pairs are represented at the right by a diamond ("same") and a square ("different").

The latencies here exhibited large variation, and their pattern was not very revealing. There was a significant main effect of closure duration ( $F_{3,6} = 9.44, p < .02$ ), apparently due to the unusual curvilinear relationship of "same" (VCV-VC) latencies to closure duration; however, the corresponding triple interaction did not reach significance. The other latencies seemed to be fairly independent of closure duration. The only other significant effect was the same-different by temporal order interaction ( $F_{1,8} = 7.50, p < .03$ ). It can be seen that the order of the two stimuli in a pair had opposite effects on "same" and "different" latencies: the former were faster in VC-VCV pairs, the latter in VCV-VC pairs. Expressed differently, (correct) "same" responses were faster than (correct) "different" responses in VC-VCV pairs, but in VCV-VC pairs "different" responses were faster than "same" responses at the two intermediate closure periods. Latencies for VC-VC pairs tended to be shorter than for pairs containing a VCV utterance, with "same" latencies being shorter than "different" latencies. This pattern indicates that VC-VC pairs were perceptually more like VC-VCV pairs than like

VCV-VC pairs, that is, the nature of the first stimulus in a pair made a difference.

The pattern of errors (lower right-hand panel in Figure 3) was considerably clearer than that of the latencies. The overall error rate was 10.4 percent, with individual rates varying from 1.6 percent to 21.9 percent. Errors for all four kinds of latencies decreased over the first three closure durations, reached a minimum at 265 msec, and then increased slightly at 365 msec. In this respect, they resemble the errors for VCV-VCV pairs with different final vowels. In contrast to the latencies, fewer errors were committed on "same" trials regardless of condition, but the magnitude of the same-different effect was much longer in VC-VCV pairs than in VCV-VC pairs. VC-VC pairs showed fewer errors than pairs containing a VCV stimulus, and again the large same-different effect makes them resemble VC-VCV pairs more than VCV-VC pairs.

#### DISCUSSION

The results of Task 1 establish that the VC portions of the stimuli /ab/, /ad/ were easy to discriminate, thus eliminating a possible objection to the earlier experiment of Repp (1976). From the results of Tasks 2 and 3, it is evident that the subjects (with one possible exception in Task 2) did not exclusively rely on the explosive transitions of VCV syllables in making their judgments, in contrast to my earlier study. RTs increased little or not at all with closure duration; certainly the increases were far from the linear function with a slope of 1.0 observed by Repp (1976). The different outcome in the present study was probably due to the intermixing of VC and VCV syllables, plus perhaps the randomization of the closure duration variable. RTs depended less on closure duration in Task 3 than in Task 2, probably because the proportion of VC stimuli was higher in the former (27 percent vs. 11 percent).

The VC stimuli acted both as a reminder to attend to the VC portion only and as a control--if subjects tried to attend to the CV portion of VCV utterances, they were likely to "wait" for it even in VC stimuli and subsequently be surprised by its absence, leading to relatively long latencies. The relatively short latencies for VC-VC pairs in Task 3 indicate that the subjects were quite successful in attending to the VC portion of the stimuli. In Task 2, on the other hand, they were less successful, and VC latencies tended to be longer than VCV latencies. However, the results of the three experienced subjects in Task 2 show quite clearly that it is possible to pick out the VC portion and completely ignore the CV portion of a VCV syllable, regardless of the duration of the interval separating the two portions. This statement must be limited by the factors of (1) experience with synthetic speech and fast RTs, (2) the particular task employed, and (3) the shortest closure period used here (65 msec). The less experienced listeners presumably employed some mixture of strategies in Task 2, making their decisions sometimes on the basis of the VC transitions, sometimes on the basis of the CV transitions. Thus, their RTs may be thought of as a mixture of two RT distributions, one reflecting decisions based only on implosive transitions, the other reflecting decisions based only on explosive

transitions.

Although the present data are not sufficient to permit an analysis of actual latency distributions, there are nevertheless some problems with this interpretation. Consider the fact that in order to make a decision on the basis of the explosive transitions, the listener has to take in at least the first 50 msec of the CV portion following the closure period. In addition, he or she must make a decision, that will take on the average, at least 200 msec. These 250 msec must be added to the closure duration, in order to arrive at the minimal latency of a response that could conceivably be based on the explosive transitions alone. This limit ranges from 315 msec at the shortest closure period to a respectable 615 msec at the longest. From Figure 2, it can be seen that all average RTs of subject 2, those of subjects 3 and 7 at the two longest closures, and those of subjects 1, 5, and 6 at the longest closure, fall below those estimates and therefore can, at best, reflect a small proportion of decisions based on the CV portions alone. Nevertheless, the latencies of subjects 2 and 7 increase with closure duration (or, seen the other way, decrease as closure duration gets shorter), with especially long RTs to VC syllables. If all individual RTs that fall above the estimate minimal CV RT are omitted from the data, the increase is still observed. This seems to suggest that, in these cases at least, decisions were initiated before the CV portion occurred, but when the explosive transitions entered the listener's ear, they facilitated the ongoing decision and thus led to faster latencies. If this interpretation is correct, it suggests that the decision process does not simply start and then spin off autonomously, but that there is continuous feed-forward of information accumulated by the perceptual system during the decision period and perhaps even during the initial stages of the initiation of motor commands. This hypothesis deserves further investigation.

Of the four types of stimulus pairs in Task 3, VC-VCV and VC-VC pairs resemble the situation in Task 2. The main difference is that in Task 2 the listener has to generate his or her own VC "target," while in VC-VCV and VC-VC pairs, the VC target is presented immediately prior to each VCV utterance and may be held in auditory memory. As a result, latencies did not increase with closure period in Task 3. This seems to suggest that the listeners simply ignored the CV portion of the VCV stimulus. However, VC-VC RTs were shorter than VC-VCV latencies, and the error rate decreased with closure duration and were smallest for VC-VC combinations. This indicates that, on the contrary, the CV portion did play a role. In Experiment II, a similar situation was encountered: see there for further discussion.

The results for VCV-VCV pairs present a true paradox. On the one hand, both "same" latencies and "same" errors changed little with closure duration, except for an increase in RTs between the two longest closure durations. This again suggests little dependence on the CV portion of the second stimulus. However, there was a large coperception effect of the final vowel at all closure periods. In particular, "same" RTs and accuracy were facilitated when the final vowels were identical. At the long closure times, this finding becomes rather puzzling, since the CV portion of the second stimulus entered the listener's auditory system as much as 365 msec after the

decision process had presumably been initiated. The results of the three experienced subjects in Task 3 ultimately mystify the results. It turned out that all three listeners showed a sizable copercception effect at the two longest closure periods, although they responded so fast that, in most cases, the CV portion of the second VCV stimulus had not even entered their ears! The effect persisted when only those RTs were retained for which this was strictly true. This points to an artifact, although the cause for the anomalous result has not been found. Thus, the copercception effect for the naive subjects must be viewed with caution, too.

In summary, these results do indicate that listeners can utilize the information conveyed by implosive transitions, although subsequent information (explosive transitions and final vowel) seems to influence the latency of the decision process. In view of the paradoxical nature and complexity of the results of Task 3, the second experiment followed up the simpler Task 2 supplemented by two additional tasks.

## EXPERIMENT II

Like its predecessor, this experiment was comprised of three tasks. Task 1 was aimed at determining the duration of the closure period that enables listeners to hear geminate consonants, that is, VC<sub>1</sub>-C<sub>1</sub>V instead of VC<sub>1</sub>V. Dorman et al. (1975) found that this duration substantially exceeds 90 msec, but they did not determine the precise perceptual boundary. Pickett and Decker (1960) determined such a boundary in TOPIC vs. TOP PICK by varying the closure period of /p/ by tape splicing. The boundary was around 200 msec at a normal speech rate. Fujisaki, Nakamura, and Imoto (1975) have recently reported a slightly shorter boundary for /t/ vs. /tt/ in Japanese. No data were available on voiced stops, so that the determination of the single-geminate boundary for such stimuli seemed a useful undertaking. The method was to vary the duration of the silent closure interval separating the VC and CV portions and to ask the subjects to judge whether they heard one or two consonants.

Task 2 extended the choice RT task of the previous experiment. First of all, it was thought that the inclusion of a greater proportion of VC syllables among the VCV stimuli would facilitate the focusing of attention on the VC portion alone. Second, a control condition was included in which the subjects had to rely on the explosive transitions; the implosive transitions of the stimuli were removed and replaced by steady-state vowel formants. The duration of the closure period was varied over a range well below the single-geminate boundary, and all stimuli were presented in a completely randomized sequence. It was expected that RTs to stimuli without implosive transitions would increase linearly with closure duration with a slope of 1.0, but RTs to stimuli with implosive transitions should be independent of closure duration (that is, have a slope of 0.0) if the subjects could really base their decisions on the VC portions only. This design took advantage of the observation that stimuli with and without implosive transitions are rather difficult to discriminate.



Task 3, in fact, required the subjects to make such discriminations at various closure durations up to and exceeding the single-geminate boundary. It was expected that discrimination performance would be close to chance at short closure periods and improve as closure duration was increased. The closure duration where performance reaches its asymptote was of special interest: Would it be in the region of the single-geminate boundary (Task 1) or earlier?

## METHOD

### Subjects

The subjects were the author and nine new, relatively inexperienced subjects.

### Stimuli

The basic stimuli were the same four VCV utterances as in Experiment I. In task 1, closure duration was varied between 65 and 515 msec in ten steps of 50 msec. The resulting 40 stimuli were recorded in three different randomizations. The interstimulus interval (ISI) was 3 sec.

In Task 2, closure duration was varied between 65 and 165 msec in steps of 25 msec. For each of the resulting 20 VCV stimuli, a corresponding stimulus without implosive transitions (V-CV) was synthesized by replacing the formant transitions with the steady-state frequencies appropriate for the initial vowel /a/, leaving all amplitude settings unchanged. These 40 stimuli were randomized, together with 20 VC stimuli (10 each of /ab/ and /ad/) in five blocks resulting in a total of 300 stimuli. The ISI was 3 sec.

In Task 3, VCV closure duration was varied from 65 to 265 msec in steps of 25 msec (the 240 msec interval was omitted), resulting in 32 stimuli. These stimuli were paired with their V-CV counterparts in an AXB paradigm. Each pairing occurred in all four AXB triad configurations (AAB, ABB, BAA, BBA), resulting in 128 stimulus triads that were recorded with ISIs of 1 sec within triads and 3 sec between triads.

### Procedure

All three tasks were administered in a single session lasting about 90 minutes. Their sequence was the same for all subjects. The equipment used was the same as in the previous experiment.

Task 1 was preceded by four examples of VCV utterances (the four basic stimuli with 65 msec closure periods) and four examples of VC<sup>1</sup>-C<sub>1</sub>V utterances (the same stimuli with 515 msec closure periods). All subjects agreed that the latter had two identical consonants in the middle, while the former had only one. The subjects were then told that they would hear a series of intermediate cases and that they should write down a 1 when they heard one consonant, and a 2 when they heard two.

Task 2 was preceded by a short practice list of 20 VC syllables. The response keys were labeled AB and AD, and the subjects were instructed to press the appropriate key as fast as possible. With respect to the experimental series that followed, they were told to concentrate on the initial portion of the stimuli, /ab/ or /ad/, and to ignore what follows. They were told that frequently there would be no vowel at the end and that these VC trials had the purpose of reminding them of their "targets" and of revealing whether they followed the instructions. The subjects were not informed about the presence of stimuli without implosive transitions. Only a single subject commented spontaneously afterwards that he had heard utterances that sounded like /a-bi/, etc. Most subjects seemed to remain unaware of the presence of these stimuli.

Task 3 was preceded by a practice series of 16 AXB triads with 265 msec closure periods. The subjects were instructed to write down A whenever the second utterance was equal to the first, and B when it was equal to the last utterance. The nature of the difference to listen for was illustrated vocally by the author, and all subjects agreed after listening to the practice trials that they knew where and what the relevant difference was.

## RESULTS

### Task 1

Single-geminate boundaries were determined for each subject by linear interpolation. The mean boundary of the ten subjects was 213 msec, with a standard deviation of 28 msec. The range observed was from 178 to 244 msec. There was no significant variation as a function of either consonant or final vowel.

### Task 2

For each individual stimulus, the median latency of all replications was calculated. The average median RTs are shown in Figure 4, together with the regression lines for the VCV and V-CV stimuli. It can be seen that they conformed quite well to the predictions. RTs to V-CV stimuli increased with closure duration, and although the increase was not quite linear, the regression line had a slope close to 1.0 as expected. RTs to VCV stimuli, on the other hand, hardly increased with closure duration; the slope of the regression line was 0.16. The statistical analysis confirmed this impression. RTs to V-CV stimuli were significantly slower than RTs to VCV stimuli ( $F_{1,9} = 181.59$ ,  $p < .01$ ), and in addition to a significant main effect of closure duration ( $F_{4,36} = 10.85$ ,  $p < .01$ ), there was a significant interaction of closure duration with presence vs. absence of implosive transitions ( $F_{4,36} = 9.72$ ,  $p < .01$ ). When the VCV RTs were analyzed separately, the main effect of closure duration did not reach significance ( $F_{4,36} = 2.26$ ,  $p > .05$ ).

Neither the consonant /b/ vs. /d/, nor the final vowel /ε/ vs. /i/, had any effect on VCV RTs. For V-CV RTs, there was a significant interaction of the two factors ( $F_{1,9} = 10.05$ ,  $p < .05$ ) that was entirely due to longer latencies for /a-di/. It seemed that /di/ lost some of its quality and

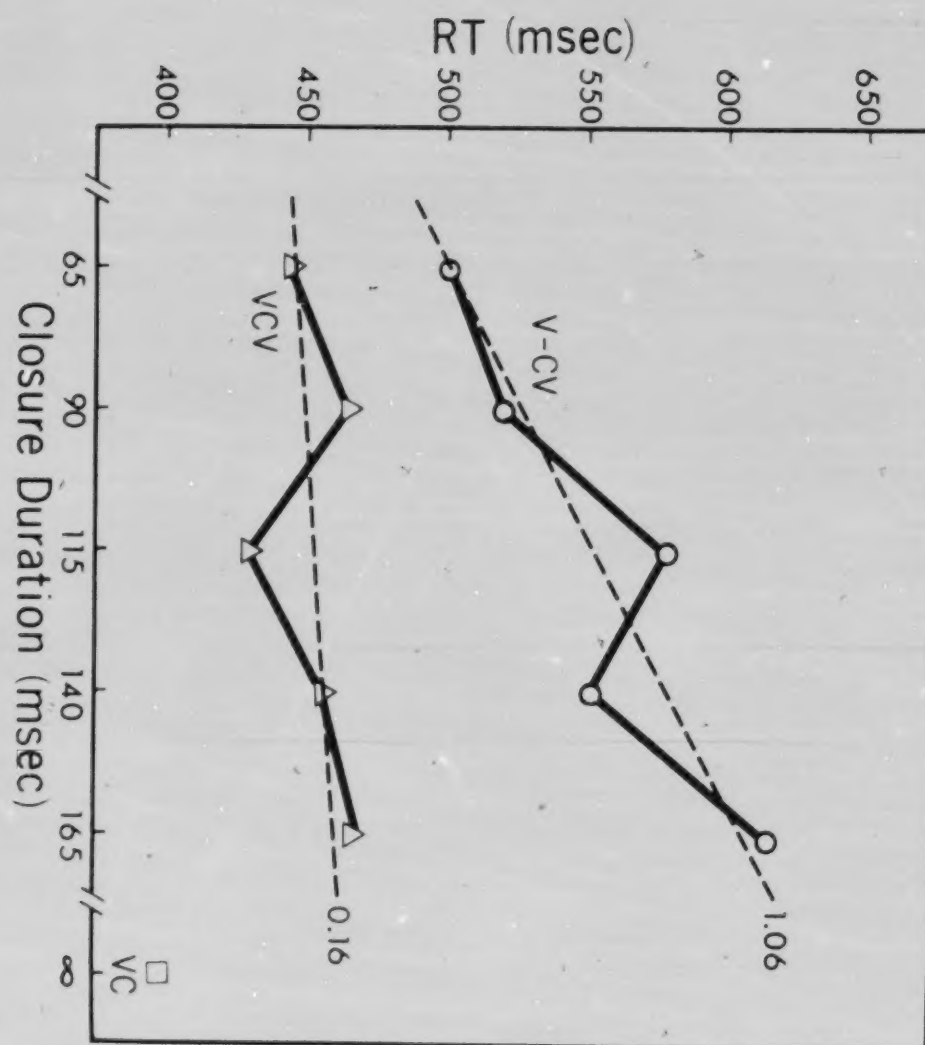


Figure 4: VCV and V-CV choice-reaction times (RTs) as a function of closure duration. The dashed lines are the corresponding regression lines, with the slopes indicated by the numbers at the right. VC reaction times are shown in the lower right-hand corner.

resembled a glottal stop when the implosive transitions were not present.

RTs to VC syllables were faster than to VCV syllables ( $F_{1,9} = 12.40$ ,  $p < .010$ , as shown in Figure 4. The average latencies for the practice VC syllables were faster by another 66 msec ( $F_{1,9} = 28.14$ ,  $p < .01$ ).

The average error rate was 7.8 percent, with individual rates varying from 0.7 percent to 28.7 percent. An additional 1.5 percent of the latencies were considered anticipations and excluded. There were many more errors for V-CV utterances (14.7 percent) than for VCV (4.6 percent) or VC (4.1 percent) utterances. There was little change in the error pattern with closure duration, in contrast to the V-CV latencies. Stimuli containing a /d/ had higher error rates than stimuli containing a /b/ in V-CV (19.8 percent vs. 9.6 percent), VCV (7.0 percent vs. 2.2 percent), and VC (4.8 percent vs. 3.4 percent) stimuli. The error rate for the practice VC stimuli was 4.5 percent.

### Task 3

One subject performed at chance level in this task; her data were excluded. The average results of the remaining 9 subjects are shown in Figure 5. Error percentages are shown as a function of closure duration, separately for /b/ and /d/ stimuli. It can be seen that errors decreased as closure duration increased and that more errors were made with /b/ than with /d/ stimuli. Somewhat against the expectation, performance was better than chance at all closure intervals, except perhaps for /b/ stimuli with the shortest closure. The shallow slope of the discrimination function was also unexpected, as was the sudden increase in errors around 200 msec closure duration.

The difference between /b/ and /d/ stimuli is best explained as an artifact due to the particular synthetic stimuli used. As pointed out earlier, /ad/ had a somewhat strident offset (probably due to the high final frequency of the third formant) that, in VCV utterances, could be detected as an acoustic residue accompanying the stimulus. It disappeared with removal of the implosive transitions and thus provided an additional discrimination cue. It was also noted earlier that the phonetic quality of /d/ (in /adi/, at least) suffered somewhat when the implosive transitions were removed. The /b/ stimuli were free from these artifacts and therefore represent a cleaner assessment of the difficulty of the task.

### DISCUSSION

The average single-geminate boundary of 213 msec found in Task 1 agrees well with the results of Pickett and Decker (1960) for voiceless labial stops. It is slightly longer than the boundary for voiceless alveolar stops in Japanese (Fujisaki et al, 1975). Both Pickett and Decker and Fujisaki et al. demonstrated that the single geminate boundary is not very stable; but varies considerably with context, particularly with the syllabic rate of a speech precursor. Certainly, this must also apply to the present boundary, so that it should not be understood as a fixed value.



The 213 msec boundary confirms one possible objection to Repp's (1976) experiment that used closure durations up to 215 msec: the subjects may not have perceived geminates in these stimuli. However, the present experiment demonstrates that it is possible to make decisions about the VC portions of VCV utterances with closure durations that are clearly on the single-consonant side of the boundary. Therefore, this cannot have been the limiting factor in Repp's (1976) experiment.

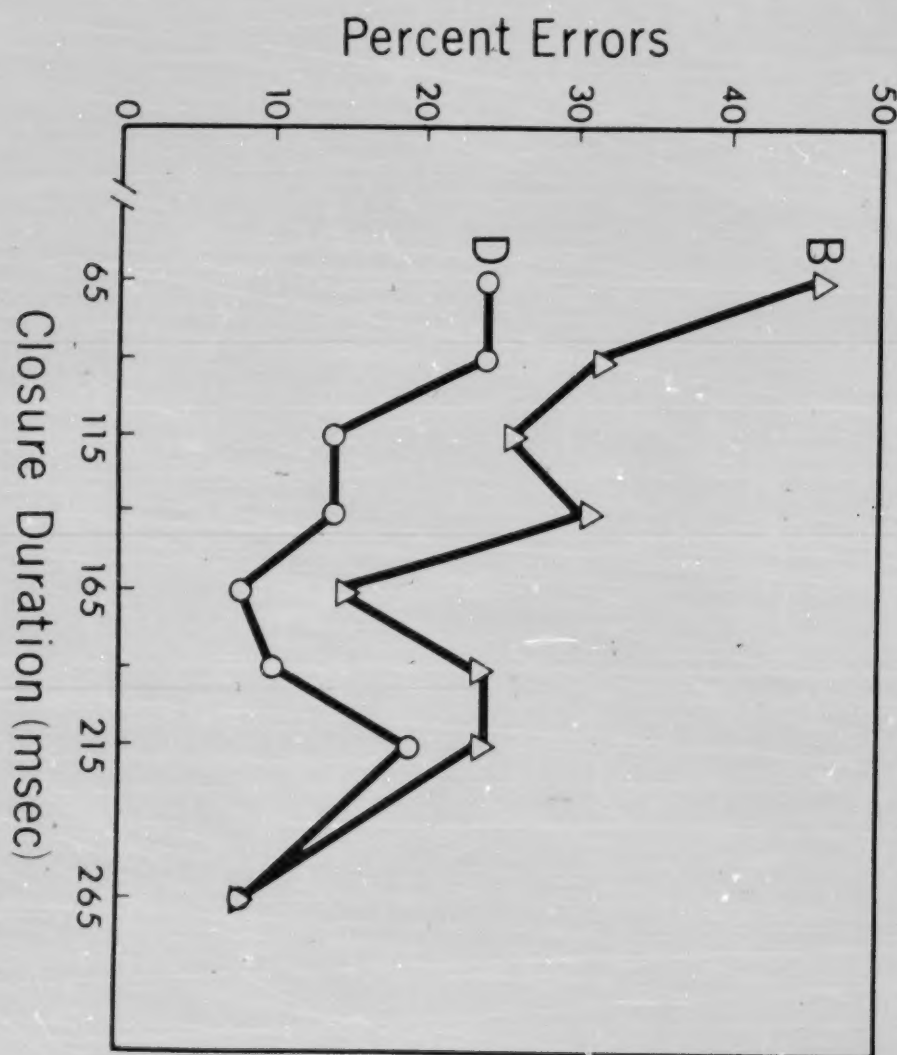


Figure 5: Percentage of AXB discrimination errors as a function of closure duration shown separately for stimuli containing a /b/ (B) and stimuli containing a /d/ (D).

Consider now the results of the discrimination experiment (Task 3). They demonstrate that, although all except the three longest closure durations were well within the single-consonant region, the presence versus absence of implosive transitions could be detected with better than chance success. Subjectively, it was nevertheless a very difficult task, as shown by the subjects' comments and my own experience as a subject. In fact, the relatively low error rates were quite surprising, since I felt I was guessing most of the time, and so did the other subjects. This suggests that the correct discriminations were based on rather subtle auditory distinctions.

The decrease of errors with closure duration took a different course than had been expected. It was thought that performance would remain poor up to about the single-geminate boundary and then rapidly rise to an asymptote. Instead, the error function decreased steadily between 65 and 165 msec closure duration. Surprisingly, there was a sudden rise in error probability precisely in the region of the single-geminate boundary. It is possible that this rise represents a transition from one kind of cue to another: at short closure durations, a subtle auditory cue is used to discriminate two phonetically equivalent utterances (V-CV vs. VCV), and this cue becomes more and more effective as the closure durations approach the single-geminate boundary. At this point, the listeners begin to utilize a phonetic cue (V-CV vs. VC-CV, that is, single vs. geminate). At first, however, this cue is relatively unreliable and therefore causes a temporary rise in error rates. The initial closure duration itself may lead the listener to focus on the phonetic level, with congruent neglect of the auditory memory representations of the stimuli. It also may be that V-CV stimuli sometimes are based on VC-CV when the closure duration is around 200 msec.

The results of Task 3 suggest that the region of chance discrimination coincides with the region in which implosive transitions that conflict with the explosive transitions are not perceived as a separate event, that is, where VC<sub>2</sub>V is heard instead of VC<sub>1</sub>-C<sub>2</sub>V. For synthetic syllables, this region extends up to about 50-80 msec (Dorman et al., 1975). This was determined in identification tasks; the corresponding discrimination experiment remains to be done. The present Task 3 differs from the conflicting transitions paradigm of Dorman et al in that all transitions (including that of the first formant) were eliminated and the higher formants were neutral and not phonetically conflicting with the explosive transitions. A study that directly compares the discriminability of the two kinds of changes at short closure durations is now in progress.

The results of Task 2 seem to provide a fairly clear demonstration that listeners can make decisions on the basis of implosive transitions in VCV utterances, even if they are not heard as a separate event. Utterances with and without implosive transitions sounded so similar that the presence of two kinds of stimuli was not noticed by most subjects. Nevertheless, the latencies for the two types of utterances differed radically: those for stimuli without implosive transitions increased linearly with closure duration because the listeners had to rely (although they may not have been fully aware of it) on the explosive transitions that occurred progressively later in the stimulus as closure duration increased. In utterances with implosive transitions, on the other hand, the latencies hardly increased with closure

duration and were considerably faster, even at the shortest closure duration used. Thus, the listeners clearly did not wait for the explosive transitions to occur before they made their decisions.

It is of some interest that the two regression lines in Figure 4 have identical intercepts (433 and 435 msec, respectively). If it is legitimate to interpolate backwards linearly, this suggests that at zero closure duration, where the offset of the impulsive transitions coincides with the onset of the explosive transitions, decisions on the basis of the explosive transitions would be just as fast as decisions based on the impulsive transitions--actually a little faster because the listener must first hear a portion of the explosive transitions, which here is included in the latencies. It is possible, however, that the pattern will change at short closure durations; an experiment is in progress to investigate this issue further.

The seemingly straightforward conclusion that the subjects relied only on the impulsive transitions in VCV stimuli is challenged by the fact that RTs to VC syllables were markedly faster. This result is in contrast to that of Task 2 in Experiment I where VC latencies were generally slower than VCV latencies. These conflicting results may have something to do with the frequency of occurrence of the VC stimuli: In Experiment I, only one-ninth of the stimuli were VCs, but in Experiment II they amounted to one-third. In order for stimulus frequency to have this selective effect, however, it must be the case that the CV portion of VCV stimuli did affect RT. Since, in Experiment II, this effect seemed to inhibit rather than to facilitate (as in Task 3 of Experiment I), the question arises why VCV RTs did not decrease as closure duration increased. After all, the latencies for VCVs must eventually approach those for VCs, given that the closure period is sufficiently long.

Thus, we again must consider the probability that the VCV RTs represent a mixture of two latency distributions: a faster VC distribution and a slower CV distribution. As closure duration increases, the mean of the CV distribution increases, but simultaneously the proportion of latencies from this distribution decreases because the longer interval facilitates reactions to the VC portion alone. If these two factors are in approximate balance, a flat latency function may be obtained for VCVs. If this model is correct, the results indicate that the VC portions can be picked out on a certain proportion of the trials, and that this proportion increases with closure duration.

This argument can be further supported by assuming that the average probability of a VC reaction,  $p(\text{VC})$ , can be estimated from the results of Task 3, the discrimination experiment. Table 1 illustrates the procedure. The first two rows show the percentages of correct discriminations and errors at the five shortest closure durations in Task 3. Let us assume that failures to discriminate indicate that the subject "missed" the VC portion and "registered" the CV portion only (CV reaction), so that the VCV stimulus could not be discriminated from the V-CV stimulus that always receives a CV reaction. Since, on trials where this occurs, the subject will make a random guess, a proportion of the correct responses equal to the error percentage will represent correct guesses following failures to register the impulsive

transitions. This correction for guessing is implemented in the next two rows of Table 1; the estimated proportions of CV reactions,  $p(CV) = 1 - p(VC)$ , equal twice the observed error rate.<sup>2</sup> The next two rows in Table 1 give the observed average V-CV and VCV RTs (cf. Figure 4). The observed average RT to VC stimuli was  $l(VC) = 395$  msec. The row "estimated VCV RTs" gives the values calculated from the formula:  $\hat{l}(VCV) = p(VC)l(VC) + p(CV)l(V-CV)$ , that is, the estimated VCV RT,  $\hat{l}(VCV)$ , is assumed to be a weighted average of the observed VC and V-CV latencies with the estimated proportions of VC and CV reactions as weights. It can be seen that the prediction leads indeed to a flat function that does not vary with closure duration, except for a slight dip at the end. As a rough approximation to the obtained results it is quite satisfactory, as shown by the deviations in the last row of Table 1. Unfortunately, the present data do not permit a more detailed test of this model.

An alternative test of the hypothesis would be to analyze only those VCV RTs that could not conceivably have been based on the CV portion. Again, we may argue that at least 250 msec will be needed from the onset of the CV portion to the occurrence of a CV reaction so that the minimal RTs to the CV portion range from 315 to 415 msec over the five closure durations. All latencies shorter than these limiting values should reflect only VC reactions and therefore should be equal to or faster (because of the cut-off of the upper tail of the latency distribution) than latencies to VC stimuli. Unfortunately, no subject had sufficiently short latencies for this analysis to be carried out, so that the results remain merely suggestive of a mixture of two latency distributions. Moreover, the analysis illustrated in Table 1 cannot easily be applied to Task 2 of Experiment I because VC latencies tended to be longer than VCV latencies there. Since a tendency to wait for the explosive transitions in VCV syllables is likely to slow down latencies to VC stimuli, the actual situation is probably more complicated than suggested in the last two paragraphs.

There were at least two subjects in Experiment II to whom the preceding discussion does not apply; they were the author and the relatively most experienced listener of the remaining subjects. Their VC latencies were not faster than their VCV latencies, and the latter were reasonably constant with closure duration, suggesting that these subjects based their decisions exclu-

<sup>2</sup>Strictly speaking, this is true only for AXB trials composed of two V-CV stimuli and one VCV stimulus. There  $p(C) = p(VC) + 1/2p(CV)$  and  $p(E) = 1/2p(CV)$ , so that  $p(VC) = 2p(E)$ , where  $p(C)$  and  $p(E)$  are the observed proportions of correct responses and errors, respectively, and  $p(VC)$  is the probability of a VC reaction. On trials where two VCV stimuli are paired with one V-CV stimulus, the probabilities are slightly more complex:

$$p(C) = p(VC)^2 + 1/2 p(CV)^2 + 1/2 p(VC)p(CV)$$

$$p(E) = 1/2 p(CV)^2 + (3/2)p(VC)p(CV)$$

This complication has been neglected in Table 1 which serves only illustrative purposes.



sively on the VC portion of the VCV utterances. The difference between VCV and VC latencies was most pronounced for the poorest subjects, but even those subjects showed evidence that at last a proportion of the responses were initiated by the implosive transitions. Future experiments will have to consider a methodological issue as well: while the many interspersed VC stimuli were intended to force the subjects to focus on the VC portions of the VCV stimuli, it may be that this effect was partially cancelled by the presence of the V-CV stimuli that forced the subjects to wait for the CV portion (although they were usually not aware of this).

#### GENERAL DISCUSSION

In contrast to my earlier two experiments (Repp, 1975, 1976), the present studies demonstrate that listeners can make decisions on the basis of the implosive transitions in VCV utterances, although few listeners seem to be able to do so consistently. We need to ask why this task is such a difficult one for inexperienced listeners.

First of all, let us consider the possibility that backward recognition masking is involved (Massaro, 1975). Since the critical closure duration at the single-geminate boundary is of approximately the same magnitude as the interstimulus interval (ISI) at which typical backward recognition masking functions reach their asymptote (Massaro, 1972, 1974), it seems possible that the implosive transitions are masked by the explosive transitions (Massaro, 1975) at closures shorter than 200 msec or so. However, the results of Dorman et al (1975) make this seem very unlikely. Dorman et al showed that only 50-80 msec of silence are necessary to perceive implosive transitions that conflict with the explosive transitions, that is, that cue a different place of articulation. Since the range of backward recognition masking supposedly reflects the time to read out information from preperceptual storage, there is no reason why masking should last longer for compatible than for incompatible transitions. The masking hypothesis therefore applies only to closure periods shorter than those used in the present experiments, and even there it may not be the correct explanation (cf. Dorman et al, 1975).

The single-geminate boundary then most likely reflects the operation of a phonetic recoding mechanism. At the auditory level, implosive and explosive transitions for the same consonant have little in common. This is suggested by the absence of cross adaptation effects between initial and final stops (Ades, 1974) and by psychophysical results on transition perception (Klatt and Shattuck, 1975). At the phonetic level, on the other hand, implosive and explosive transitions of the same consonant converge upon the same phonemic unit because they are interpreted phonetically in relation to their vocalic contexts. Like a variety of other auditory features that are spread out in time, they will be integrated into a single phonemic representation (that is, coperceived) as long as they occur within a certain interval of time. In this view, the closure duration at the single-geminate boundary plus the actual durations of the implosive and explosive transitions reflects the time window over which the phonetic recoding mechanism integrates; it seems to be on the order of 300 msec. This seems to be long enough to accommodate all major coarticulation effects in speech, so that it may

represent a general upper limit of phonetic integration. Being a higher-level mechanism, it is not surprising that the phonetic integrator is flexible and sensitive to context. A change in its time window with a change in the rate of speech is highly adaptive and almost a necessity (Pickett and Decker, 1960; Fujisaki et al, 1975). It is likely that implicit knowledge about articulatory constraints is also utilized at this level (cf. Dorman et al, 1975).

It is intuitively plausible that inexperienced listeners will have a strong tendency to rely on their phonetic codes when making judgments about speech. After all, this is the natural way of dealing with the speech message in everyday life where the phonetic code is supplemented by prosodic characteristics that are available in auditory memory. Results such as Repp's (1976) probably reflect the exclusive use of phonetic representations. The "coperception effect" in the same-different task may arise from the listener's waiting for the complete phonetic code (of the second stimulus in a VCV pair) to be established and a subsequent attempt to match the two phonetic codes holistically, that is, as words (or perhaps as CV syllables, if the redundant initial vowel is ignored). If this match fails, the two consonant phonemes are extracted and compared. As a result, "same" latencies will be shortest for completely identical stimuli, while "different" latencies will not show a consistent coperception effect. Pisoni and Tash (1974) thought that the holistic matching operation may take place at the auditory level, but because of the limitations of auditory storage (see below) it is better attributed to the phonetic level.

On the other hand, consider the practiced listeners who, in the choice RT tasks, apparently were able to selectively attend to the VC portion only. Their success can be explained in two ways: either they were able to treat the phonetic code more analytically, thus bypassing the holistic matching stage, or they were able to perform a selective matching operation at the auditory level. There is considerable evidence in the literature that stop consonants are poorly retained in auditory memory (Fujisaki and Kawashima, 1970; Crowder, 1971, 1972, 1973; Pisoni, 1971; Cole, 1972; Darwin and Baddeley, 1974). Nevertheless, the possibility of auditory matching cannot be completely ruled out. It was mentioned earlier that the auditory quality of the VC offsets was different (mellow for /ab/, strident for /ad/)--which was not so much an artifact as an inherent property of the synthetic speech used--and the experienced listeners may have focused on this difference. It seems, however, that this rather subtle auditory discrimination should have been associated with longer latencies than obtained from the experienced subjects. Moreover, the auditory difference between /ab/ and /ad/ seemed to be largely obliterated in VCVs with short closure durations.

The more plausible alternative is therefore that the experienced subjects--and the less experienced subjects with varying degrees of success--relied on a more analytic use of their internal phonetic stimulus representations. There are some reasons why it may be difficult for the average listener to respond immediately to the implosive transitions: in natural speech, the closure period is either followed by explosive transitions (or a similar event signaling continuation of the speech signal) or it is terminated by a release burst, that was absent in the present VC stimuli. The

phonetic recoding mechanism probably maps the implosive transitions immediately into a phonetic buffer but assigns the corresponding phoneme a temporary or "weak" status until some further confirmation is obtained from the speech signal, either in form of a release or of explosive transitions. The implosive transitions are only a partial cue (though a sufficient one) for a final or medial stop consonant, and the phonetic processor "knows" this. The phonetic representation therefore will be complete only when the integration period of the recoding mechanism has moved beyond the implosive transitions and thus has encountered either further information or silence (in the present VC stimuli). The skill of experienced listeners may thus consist in making fast decisions on the basis of "weak" phonemes based on partial cues. If, in listening to a VCV utterance, no decision is initiated before the explosive transitions occur, the "weak" phoneme will be replaced by a "strong" phoneme based on the joint cues of implosive and explosive transitions, as long as the two cues occur within the phonetic integration period. This may be thought of as a kind of "phonetic backward masking," the strong phoneme replacing the weak one that is then no longer available. Inexperienced listeners may not initiate their decisions early enough on a proportion of trials, so that, on these trials, their decision will be based on the CV portion.

Another factor that may play a role in this task is the possibility of a hierarchical structure in the phonetic representation that tends to group consonants with the following vowel rather than with the preceding one (MacKay, 1972). If such a hierarchical structure is imposed while the phonetic code is being constructed, there would be a natural tendency to wait for the CV portion before making a decision. Again, the task to respond to the VC portion will be an unusual one that requires practice and facilitating context, such as many VC stimuli in a list. In effect, it requires the listener to modify his syllabic recoding scheme.

Thus, there are various factors that make it difficult for the average listener to single out a portion of the speech signal for immediate responding. However, the present results show that, under favorable circumstances, listeners are able to utilize the information conveyed by implosive transitions even when they are not perceived as a separate phonemic event and despite their relatively weak cue value.

#### REFERENCES

- Abbs, M. H. (1971) A study of cues for the identification of voiced stop consonants in intervocalic contexts. Unpublished Ph.D. dissertation, University of Wisconsin.
- Ades, A. E. (1974) How phonetic is selective adaptation? Experiments on syllable position and vowel environment. *Percept. Psychophys.* 16, 61-66.
- Bock, R. D. (1975) Multivariate Statistical Methods in Behavioral Research. (New York: McGraw-Hill).
- Cole, R. A. (1972) Different memory functions for consonants and vowels. *Cog. Psychol.* 3, 377-383.
- Crowder, R. G. (1971) The sound of vowels and consonants in immediate memory. *J. Verbal Learn. Verbal Behav.* 10, 587-596.

- Crowder, R. G. (1972) Visual and auditory memory. In Language by Ear and by Eye, ed. by J. G. Kavanagh and I. G. Mattingly. (Cambridge: MIT Press), pp. 251-276.
- Crowder, R. G. (1973) Representation of speech sounds in precategorical acoustic storage. J. Exp. Psychol. 98, 14-24.
- Darwin, C. J. and A. D. Baddeley. (1974) Acoustic memory and the perception of speech. Cog. Psychol. 6, 41-60.
- Dorman, M. F., L. J. Raphael, A. M. Liberman, and B. H. Repp. (1975) Some maskinglike phenomena in speech perception. Haskins Laboratories Status Report on Speech Research SR-42/43, 265-276.
- Fujimura, O. (1975) A look into the effects of context--Some articulatory and perceptual findings. Paper presented at the 8th International Congress of Phonetic Science, Leeds, England.
- Fujisaki, H., and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute, Faculty of Engineering, (Tokyo: University of Tokyo), 29, 207-214.
- Fujisaki, H., K. Nakamura, and T. Imoto. (1975) Auditory perception of duration of speech and non-speech stimuli. In Auditory Analysis and Perception of Speech, ed. by G. Fant and M. A. A. Tatham, (London: Academic Press), pp. 197-220.
- Klatt, D. H. and S. R. Shattuck. (1975) Perception of brief stimuli that resemble rapid formant transitions. In Auditory Analysis and the Perception of Speech, ed. by G. Fant and M. A. A. Tatham. (London: Academic Press), pp. 293-302.
- MacKay, D. G. (1972) The structure of words and syllables: Evidence from errors in speech. Cog. Psychol. 3, 210-227.
- Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.
- Massaro, D. W. (1974) Perceptual units in speech recognition. J. Exp. Psychol. 102, 199-203.
- Massaro, D. W. (1975) Preperceptual images, processing time, and perceptual units in speech perception. In Understanding Language. An Information-Processing Analysis of Speech Perception, Reading, and Psycholinguistics, ed. by D. W. Massaro. (New York: Academic Press), pp. 125-150.
- Pickett, J. M. and L. R. Decker. (1960) Time factors in perception of a double consonant. Lang. Speech 3, 11-17.
- Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Supplement to Haskins Laboratories Status Report on Speech Research, November.
- Pisoni, D. B., and Tash, J. (1974) "Same-different" reaction times to consonants, vowels, and syllables. In Research on Speech Perception, Progress Report No. 1. (Department of Psychology, Indiana University).
- Repp, B. H. (1975) "Coperception": A preliminary study. Haskins Laboratories Status Report on Speech Research SR-42/43, 147-157.
- Repp, B. H. (1976) Coperception: Two further preliminary studies. Haskins Laboratories Status Report on Speech Research SR-45/46, 141-152.
- Wood, C. C. and R. S. Day. (1975) Failure of selective attention to phonetic segments in consonant-vowel syllables. Percept. Psychophys. 17, 346-350.



# What Can /w/, /l/, /y/ Tell Us About Categorical Perception?\*

Lyn Frazier†

## ABSTRACT

A three-formant pattern was synthesized in such a way that systematically increasing the frequency of the second formant (F<sub>2</sub>) yielded a continuum that was perceived as /w/ at the lower F<sub>2</sub> frequencies, /l/ at the intermediate F<sub>2</sub> frequencies, and /y/ at the higher F<sub>2</sub> frequencies. Mirror-image stimuli were also constructed so that /w/, /l/, and /y/ occurred in both syllable-initial and syllable-final position. Listeners were asked to discriminate between neighboring stimuli in both a long and a short interstimulus interval (ISI) condition.

Results indicate that both sets of stimuli were perceived categorically in both ISI conditions. These results suggest that the transience of acoustic information is not a crucial determinant of mode of perception. The results also illustrate a problem with the scoring method traditionally used in categorical perception studies.

## INTRODUCTION

One of the much-studied but little-understood curiosities of human perception is a phenomenon psychologists have called categorical perception. In this study, a series of experiments have been conducted on the perception of /w/, /l/, /y/ in order to further our understanding of this phenomenon.

Specifically, these experiments were designed to investigate three questions. The first question concerns the feasibility of isolating a property, or set of properties, that determine how any particular speech sound will be perceived. The second question concerns the correctness of the explanation of categorical perception proposed by Fujisaki and Kawashima (1970). The third question explores the possibility of an alternative, "masking," explanation of categorical perception.

---

\*Submitted in partial fulfillment of the general examination requirement, University of Connecticut, Spring, 1976.

†University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

Before these questions can be examined in detail, it is necessary to review some of the previous research on categorical perception.

#### BACKGROUND

Research on speech perception has provided evidence that synthetic stop consonants, unlike synthetic steady-state vowels, are perceived categorically (Liberman, Harris, Hoffman, and Griffith, 1957). That is, listeners find it difficult to discriminate between two tokens of stop consonants drawn from the same phonetic category even though the objective difference between these tokens is equal to the easily discriminated difference between the two stop consonants drawn from different phonetic categories. The perception of synthetic steady-state vowels, on the other hand, is typically "continuous" (Fry, Abrams, Eimas, and Liberman, 1962). Listeners can discriminate many more intraphonemic differences when listening to steady-state vowels than when listening to stop consonants.

This difference in perception between stop consonants and steady-state vowels was originally attributed to the different manners in which these sounds are produced: the discrete gestures involved in the articulation of stop consonants were assumed to yield discrete perceptual categories, whereas the more variable gestures in the articulation of vowels were assumed to yield more variable perceptual categories (Liberman, 1957). In an elaboration of the motor theory, Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) proposed that in "the production system there exist neural signals standing in one-to-one correspondence with the various segments of the language. They then argued that, as the neural signals pass through the articulatory system, the phonetic segments underlying speech are encoded (restructured) to various degrees. Stop consonants were thought to undergo the greatest degree of acoustic restructuring and were therefore the most highly encoded sounds, while steady-state vowels were the least encoded. This account developed as part of a larger theory that postulated a special left hemisphere speech processor. Since consonants were highly encoded, it was thought that their perception required a specialized perception mechanism. The relatively unencoded steady-state vowels were more similar to nonspeech sounds and therefore did not require a processor specialized for speech, but could be perceived by more general auditory mechanisms. This theory raises the possibility that categorical perception might be an unique component of the specialized speech processor.

A serious problem for the motor theory, as for any theory that attempts to explain the difference in the perception of stop consonants and steady-state vowels by reference to production mechanisms, is the fact that some nonspeech sounds are also perceived categorically (Miller, 1974; Rosner and Cutting, in press). If categorical perception is due to some aspect of the speech production system, it should be confined to speech sounds.

Fujisaki and Kawashima (1969, 1970) articulated a model of the discrimination process that attributes the difference between categorical and continuous perception to a differential use of auditory short-term memory. According to this model, auditory short-term memory stores information about

the acoustic parameters of stimuli, and thus enables a listener to make a comparative judgment about the acoustic identity of two stimuli. Auditory memory is assumed to contain an analog of the original acoustic stimulus. This analog is subject to rapid decay and may be displaced if another sound arrives before decay is complete. It is further assumed that auditory memory is relatively strong for vowels, but relatively weak for consonants, perhaps due to the greater duration and intensity of vowels (Lane, 1965) and to the brief transient nature of the acoustic cues for stop consonants (Stevens, 1968).

In the Fujisaki and Kawashima model, auditory short-term memory is contrasted with phonemic short-term memory, which can only store discrete information about the phonological category of the stimulus. From this it follows that a listener who relied solely on phonemic short-term memory could discriminate only those stimuli that could be identified as belonging to different phonemic categories.

The Fujisaki and Kawashima account of categorical perception differs from earlier explanations in that it relies on the interaction of the acoustic properties of the stimulus with memory, whereas earlier explanations relied on the perception of articulatory properties of the stimulus.

Pisoni (1973) tested the Fujisaki and Kawashima model by presenting subjects with a delayed comparison recognition memory task. Pisoni hypothesized that if categorical perception were related to the differential use of auditory memory in discrimination tasks, this could be demonstrated by varying the interval between the presentation of the stimuli being compared. Because auditory memory is thought to decay more rapidly than phonemic memory, this delay should force listeners to rely on phonemic memory.

Pisoni found that varying the interstimulus interval (ISI) from the end of one stimulus to the beginning of the next stimulus in a discrimination task did not affect the discrimination of synthetic stop consonants, whereas the ISI was inversely related to listeners' ability to discriminate synthetic vowels accurately.

Pisoni's results are consistent with the Fujisaki and Kawashima explanation of categorical perception, first, because the higher discrimination of vowels at a short ISI indicates that some additional, nonphonetic, information is available for vowels, and, second, because the decay of this information (at longer ISIs) induces categorical perception of vowels. Also, the finding that longer ISIs did not affect the discrimination of consonants is predicted by the assumption that auditory memory is already relatively weak for consonants and cannot be further weakened by an increase in ISI.

Finally, it should be noted that the vast majority of experiments on categorical perception have used either consonant vowel (CV) syllables or vowels in isolation as the test stimuli. The relevance of this last observation (and the exceptions to it) will be discussed in the next section.

### EXPERIMENTAL QUESTIONS AND PREDICTIONS

In order to investigate the three questions mentioned in the introductory section, a series of identification and discrimination tests was prepared with two sets of test stimuli. One set of stimuli consisted of /w/, /l/, /y/ in syllable-initial position. The other set of stimuli consisted of /w/, /l/, /y/ in syllable-final position. Both sets of stimuli were presented in two conditions. In one condition, the test stimuli in the discrimination test were separated by a long ISI; in the other condition, the test stimuli were separated by a short ISI.

This experimental design allows us to begin investigating the questions mentioned in the introductory section. The first question asked was why some classes of sounds should be perceived categorically rather than continuously. If we are to understand categorical perception, we must know what are the crucial features of a given sound that determine how that sound will be perceived. Is it the transience of information in stop consonants which dictates that they be perceived categorically? Or is it, perhaps, the presence of a transition?

Studdert-Kennedy (1976) has suggested that categorical perception is of functional import, noting that "rapid sensory decay and transfer into a nonsensory code is probably crucial to an efficient linguistic signalling system." This interpretation suggests that the transience of information in stop consonants may not dictate that stop consonants be perceived categorically, but rather allows them to be perceived categorically. If this interpretation is correct, then the real problem becomes that of explaining the noncategorical perception of steady-state vowels. Crowder (1972) has suggested that the functional reasons for this may be that vowels carry most of the suprasegmental information in the speech signal, so that auditory information about vowels must be stored until this suprasegmental information is extracted. Alternatively, it may be that the lack of a transition in steady-state vowels (which, after all, are only an idealization of the vowels found in actual speech) prevents them from being perceived categorically. Until more is known about the properties which actually determine that a sound will be perceived categorically, these fascinating suggestions will remain in the realm of speculation.

The segments /w/, /l/, /y/ were chosen as the test sounds for the experiments reported in this study because they cannot readily be characterized as either consonants or vowels. According to Chomsky and Halle (1968), /w/, /y/ are characterized as neither consonantal nor vocalic, because they function more like consonants, even though they are acoustically more similar to vowels. On the other hand, /l/ is characterized as being both consonantal and vocalic, because it is acoustically and functionally similar to both consonants and vowels. Although these sounds do not comprise a linguistic continuum, they are all characterized by relatively long transitions (approximately twice as long as the transitions of most stop consonants), and can be synthesized along an acoustic continuum.



If we know how these sounds are perceived, we will be able to discard certain characteristics as possible determinants of mode of perception. For instance, if the perception of /w/, /l/, /y/ is found to be continuous, we can discard the presence of a transition as a crucial determinant of mode of perception. On the other hand, if the perception of /w/, /l/, /y/ is found to be categorical, we can discard transience of information as a crucial determinant of mode of perception.

The second question addressed by this study is whether the explanation of categorical perception proposed by Fujisaki and Kawashima is correct. Assuming that the ISI condition used in the present experiments is sufficiently long for decay of auditory information, the Fujisaki and Kawashima explanation of categorical perception admits only two possible outcomes: either the test stimuli are perceived categorically in both ISI conditions, or they are perceived more categorically in the long ISI condition than in the short ISI condition. If the test stimuli were found to be perceived continuously in both ISI conditions, or if they were perceived less categorically in the long ISI condition than in the short, then the proposed explanation of categorical perception would be falsified.

The third question addressed is the possibility of an alternative explanation of categorical perception consistent with the findings of previous research. As mentioned earlier, most of the previous experiments on categorical perception have used either CV or V stimuli. This raises the possibility that stop consonants do have potentially useful nonphonetic information and that their "categorical" perception is due not to any real difference in the way they are perceived, but rather to "masking" of this nonphonetic information by the following vowel. According to the "masking" hypothesis, both stop consonants and steady-state vowels would be perceived categorically at a long ISI (due to the loss or fading of auditory information) and both would be perceived continuously when presented in an unmasked (final) position at an ISI too short for loss, or fading, of auditory information, to occur.

This alternative explanation of categorical perception is consistent with previous research. To my knowledge, the only experiments on categorical perception that have been conducted with VC syllables (for example, Mattingly, Liberman, Syrdal, and Halwes, 1971; Raphael, 1972) were conducted before Pisoni (1973) had demonstrated that the exact specification of the ISI is critical in a discrimination test. According to the "masking" hypothesis, we might not expect continuous perception in a comparison task with a 1 sec. ISI (which was the standard ISI used before 1973), but only in a comparison task with a substantially shorter ISI, say of 250 msec or less. (Pisoni found that infraphonemic discrimination of vowels began to drop off after 250 msec.)

If the masking hypothesis is correct, we would expect the test stimuli in which /w/, /l/, /y/ are in syllable-final position to be perceived continuously in the short ISI condition.

## EXPERIMENTS I, II, and III

### Description of Stimuli /w/, /l/, /y/

Previous research on the acoustic cues for glides and liquids (Lisker, 1957; O'Connor, Gerstman, Liberman, Delattre, and Cooper, 1957) suggested that a three formant pattern could be synthesized in such a way that systematically increasing the steady-state onset frequency of the second formant ( $F_2$ ) would yield a continuum that would be perceived as /w/ at the lower  $F_2$  frequencies, /l/ at the intermediate  $F_2$  frequencies, and /y/ at the higher  $F_2$  frequencies.

Such a continuum was created on a Glace-Holme parallel resonance synthesizer at the University of Connecticut. Eleven tokens were selected from the continuum to be used as test stimuli. Each stimulus was 380 msec long, with a straight  $F_3$  at 2651 Hz. An initial 90 msec,  $F_1$  steady-state at 372 Hz was followed by a 290 msec steady-state at 498 Hz. The final 250 msec steady-state of the second formant, corresponding to /ε/, was also held constant for all eleven stimuli.

Only the frequency of the initial portion of the second formant was varied. Stimulus 0 had an initial 90 msec  $F_2$  steady-state at 594 Hz, followed by a 60 msec transition to the steady-state corresponding to /ε/. The other ten stimuli were created by adding ten steps of approximately 218 Hz to the value of the initial  $F_2$  steady-state of the previous stimulus.

An initial amplitude transition was also specified for each of the three formants. The  $F_1$  amplitude began at -20 dB relative to the  $F_1$  steady-state corresponding to /ε/, and rose to "0" dB. The  $F_2$  amplitude began at -29 dB and rose to -8 dB. The  $F_3$  amplitude started at -39 dB and was increased to -23 dB. The  $F_1$  and  $F_2$  amplitude transitions were 180 msec long; the  $F_3$  amplitude transition was 90 msec long. After an amplitude transition reached its peak value, it remained at that value for the duration of the stimulus.

The acoustic specifications of the stimuli, shown in schematic form in Figure 1, represent a series of accommodations between the "best" token of each phonetic category and the "best" token of the other two phonetic categories. For example, with longer transitions a slightly more convincing /wε/ could have been obtained, but only at the expense of a less convincing /lε/. The relatively long (90 msec) initial  $F_2$  steady-state was also chosen because it yielded a more convincing /lε/. Although this is much longer than the (30-60 msec)  $F_2$  steady-state used by O'Connor et al (1957), the difference is not as great as it might appear, due to the gradual  $F_2$  amplitude onset.

It should be pointed out that the  $F_2$  steady-state in Stimulus 10 is actually higher than the  $F_3$ , and it is therefore quite likely that the  $F_3$  was "heard" as the  $F_2$  in that stimulus.

After Experiment I was conducted, it was decided that 50 msec of the steady-state corresponding to /ε/ should be deleted in order to make the test stimuli more comparable to those used by Pisoni. Appropriate controls (to be

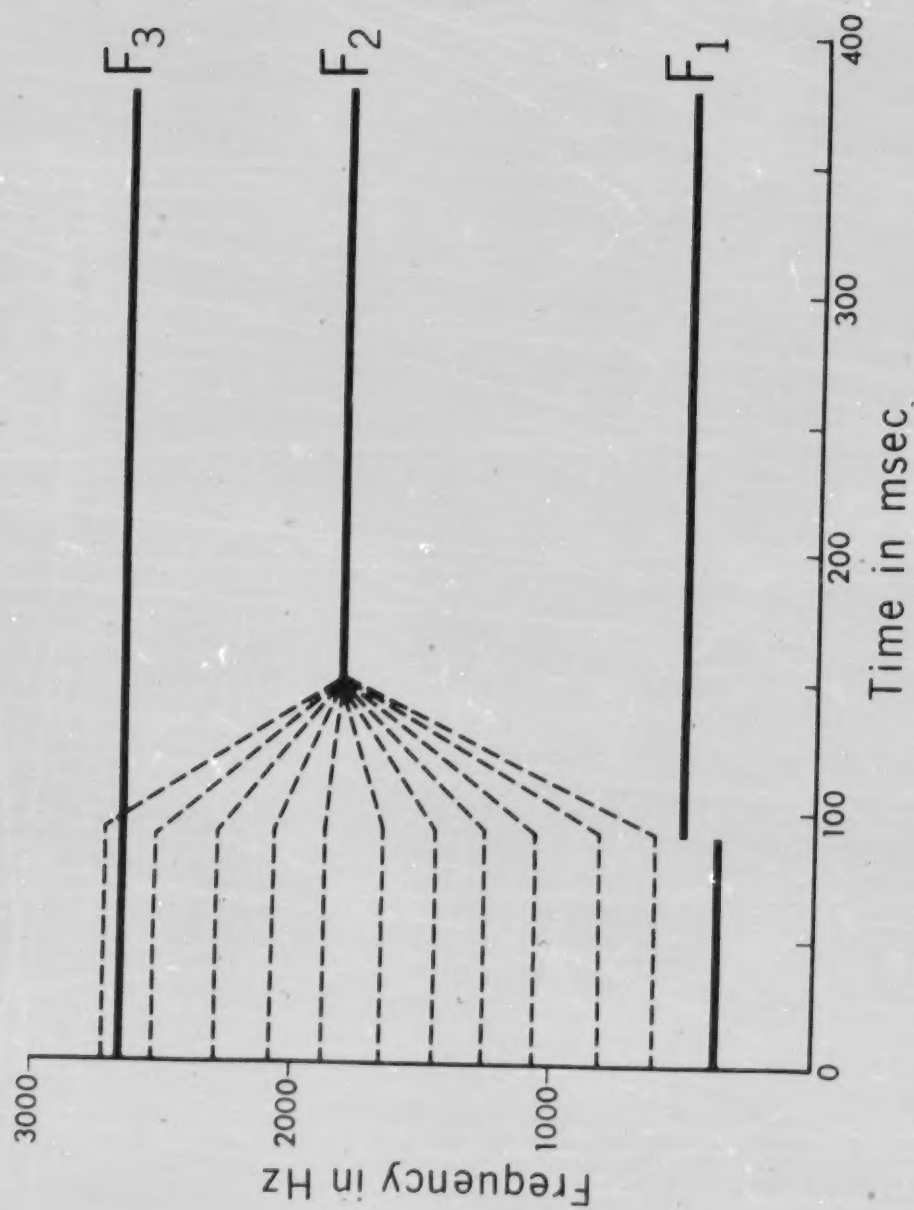


Figure 1: Stylized spectrogram of the /wɜ/-/yɛ/ stimuli.

discussed later) were run to insure that this did not influence the results.

#### EXPERIMENTS IV-VII

##### Description of Stimuli (/ew/, /el/, /ey/)

The stimuli used in Experiments IV-VII were the mirror images of the stimuli used in Experiments II and III, with one exception. The initial amplitude transitions of the original stimuli were superimposed on the mirror image stimuli because this yielded more natural sounding stimuli. In all other respects the stimuli were perfect mirror images of the original stimuli.

##### Experimental Materials

An identification test and two same-different discrimination tests were prepared. The identification test consisted of 16 different presentations of each of the 11 stimuli. Every time a stimulus was presented, it occurred twice (thus each stimulus actually occurred 32 times for a total of 16 judgments per stimulus). Each stimulus was presented twice in order to obtain consistent identification. An ISI of one sec was used to separate the two occurrences of one stimulus; an ISI of four sec was used to separate the presentations of different stimuli. The test was divided into 16 blocks, with each of the eleven stimuli in each block, in random order. These blocks were separated from each other by an eight-sec interval.

In each discrimination test, each stimulus was paired with its neighbor 16 times, and with itself 16 times. In the "long" discrimination test, an ISI of one sec was used to separate the members of a pair. In the "short" discrimination test, an ISI of 57 msec was used to separate members of a pair (because this was the shortest interval available). In both tests, an ISI of four sec was used between pairs. The stimuli were presented in random order, in 16 blocks of 11 pairs each. Blocks were separated from each other by an eight-sec interval.

##### Procedure

Subjects: Linguistically untrained undergraduates at the University of Connecticut participated in these experiments either for course credit or for \$2 per hour. All subjects were native speakers of English with apparently normal hearing.

Subjects participating in Experiments I, II, IV, and V were run in two sessions, with at least a 15-minute break between sessions. The first session consisted of an identification test followed by half of one of the discrimination tests; the second session consisted of the second half of the same discrimination test.

Subjects participating in Experiments III and VI were run in three sessions, with two breaks lasting at least 15 minutes each (occurring in the middle of each discrimination test). Half of the subjects in each experiment



took the "short" discrimination test last; the other half took the "long" discrimination test last. All subjects took the identification test first.

Subjects participating in Experiment VII were run in four sessions on two consecutive days. These sessions were identical to the sessions used in Experiments I, II, IV, and V, except that two of the subjects took the identification and "short" discrimination test for the original stimuli on the first day, and the same tests for the mirror image stimuli on the second day, with the order being reversed for the other two subjects.

Instructions for the Original Stimuli: Before the identification test, subjects were told that on each trial they would hear one of three syllables: /wε/, /lε/, /yε/. If they heard /wε/, they were to write 'W' on the answer sheets that had been provided; if they heard /lε/ they were to write 'L', and so forth.

Before the discrimination test, subjects were told that they would hear a pair of sounds. If the two members of the pair sounded the same, subjects were to write 'S',; if they sounded different, subjects were to write 'D'.

Instructions for the Mirror-Image Stimuli: The above instructions were appropriately revised for the mirror image stimuli.

#### Description of Experiments I-VII

Experiment I consisted of the identification and "long" discrimination test with the original stimuli. Experiment II consisted of the identification and "short" discrimination test with the original stimuli. Experiment III consisted of the identification and both discrimination tests done with the original stimuli.

Experiment III was run in order to ensure that the results obtained from Experiments I and II were comparable (that is, not influenced by the longer stimuli used in Experiment I or by intersubject differences).

Experiments IV, V, and VI were identical to Experiments I, II, and III, respectively, except that the mirror-image stimuli were used.

Experiment VII consisted of the identification and both discrimination tests for both sets of stimuli.

A summary of these experiments is shown in Table 1.

#### Results of Experiments I, II, and III

The data of 19 of the 20 subjects who participated in Experiments I, II, and III reveal that the perception of /w/, /l/, /y/ is categorical in the sense that discrimination was at or near chance (50 percent correct responses) within phonetic categories, and higher at category boundaries than anywhere within a phonetic category. (This point will be continued later.) The one remaining subject did not hear a single difference in the entire

TABLE 1: Summary of variables in Experiments I-VII.

## Experiment I:

Subjects: 8 paid  
 Stimuli: 380 msec  
           /we/-/ye/  
 Discrim. Test: 1 sec ISI

## Experiment II:

Subjects: 8 credit  
 Stimuli: 350 msec  
           /we/-/ye/  
 Discrim. Test: 57 msec ISI

## Experiment III:

Subjects: 4 paid  
 Stimuli: 350 msec  
           /we/-/ye/  
 Discrim. Tests: 1 sec ISI  
                   57 msec

## Experiment IV:

Subjects: 8 credit  
 Stimuli: 350 msec  
           /ew/-/ey/  
 Discrim. Test: 1 sec ISI

## Experiment V:

Subjects: 8 credit  
 Stimuli: 350 msec  
           /ew/-/ey/  
 Discrim. Test: 57 msec ISI

## Experiment VI:

Subjects: 4 paid  
 Stimuli: 350 msec  
           /ew/-/ey/  
 Discrim. Tests: 1 sec ISI  
                   57 msec

## Experiment VII:

Subjects: 4 paid  
 Stimuli: 350 msec  
           /we/-/ye/  
           /ew/-/ey/  
 Discrim. Test: 57 msec ISI

discrimination test.

Graphs of the responses of two of the subjects who participated in Experiment III are displayed in Figures 2 and 3. The identification curves shown in these graphs are typical of the identification curves of other subjects in that: (1) both show a crossover from the /l/ category to the /y/ category at stimulus 6, and (2) they indicate the variability found with respect to the crossover from the /w/ category to the /l/ category. (Three subjects did not hear a single /w/, whereas two subjects consistently labeled stimulus 2 as /w/. The remaining 15 subjects consistently labeled stimulus 0 /w/, and stimulus 2 /l/.)

The discrimination curves shown in Figures 2 and 3 were determined using the scoring method traditionally used in categorical perception studies. That is, each discrimination test was scored by combining the 16 responses from a stimulus paired with itself, with the 16 responses from a stimulus paired with its next higher numbered neighbor. The percent of correct responses was then graphed based on the (combined) 32 judgments/stimulus. This graph was then compared with the subject's identification graph.

Of particular interest in Figure 2, is the absence of a peak in the discrimination curve between the /w/ and /l/ categories in the 1-sec condition. However, a closer look at the data reveals that this subject's discrimination really was much more acute here than within categories, if we look just at the correct "different" responses.

For this reason, an alternative scoring method was also employed. This scoring method counted only the number of correct "different" responses. The percent correct "different" responses for subjects DS and KH (whose total correct responses were shown in Figures 2 and 3) are shown in Figure 4.

If we look at Figure 4, a peak in the 1-sec discrimination curve of subject DS emerges in the expected location (between stimulus 0 and stimulus 2). What then accounts for the absence of this discrimination peak in Figure 2?

In Figure 2, the traditional scoring method was used, and thus each point on the discrimination curve represents the sum of correct "same" and correct "different" responses. Because subject DH (in Figure 2) had an unusually small number of correct "same" responses (that is, a relatively large number of false "different" responses) at the /w/-/l/ category boundary, this reduced the total number of correct responses, thereby lowering his discrimination curve at the /w/-/l/ boundary.

When both the traditional and the alternative scoring method were used to tabulate the data, the 19 subjects who heard any differences in the discrimination tests had substantially more correct "different" responses (ranging from two to eight times as many) between any two categories that they consistently labeled, than they did within those categories.

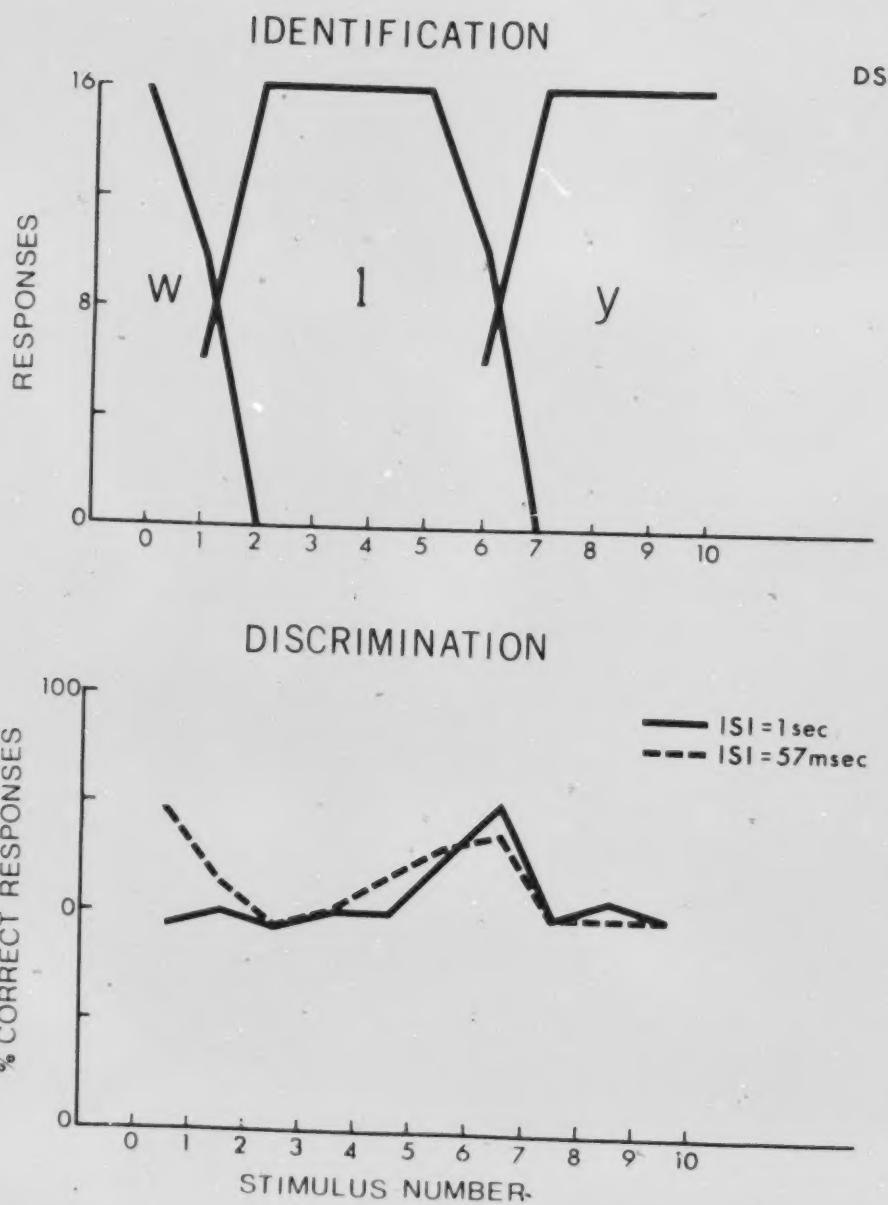


Figure 2: Identification and percent total correct discrimination responses for subject DS (Experiment III).



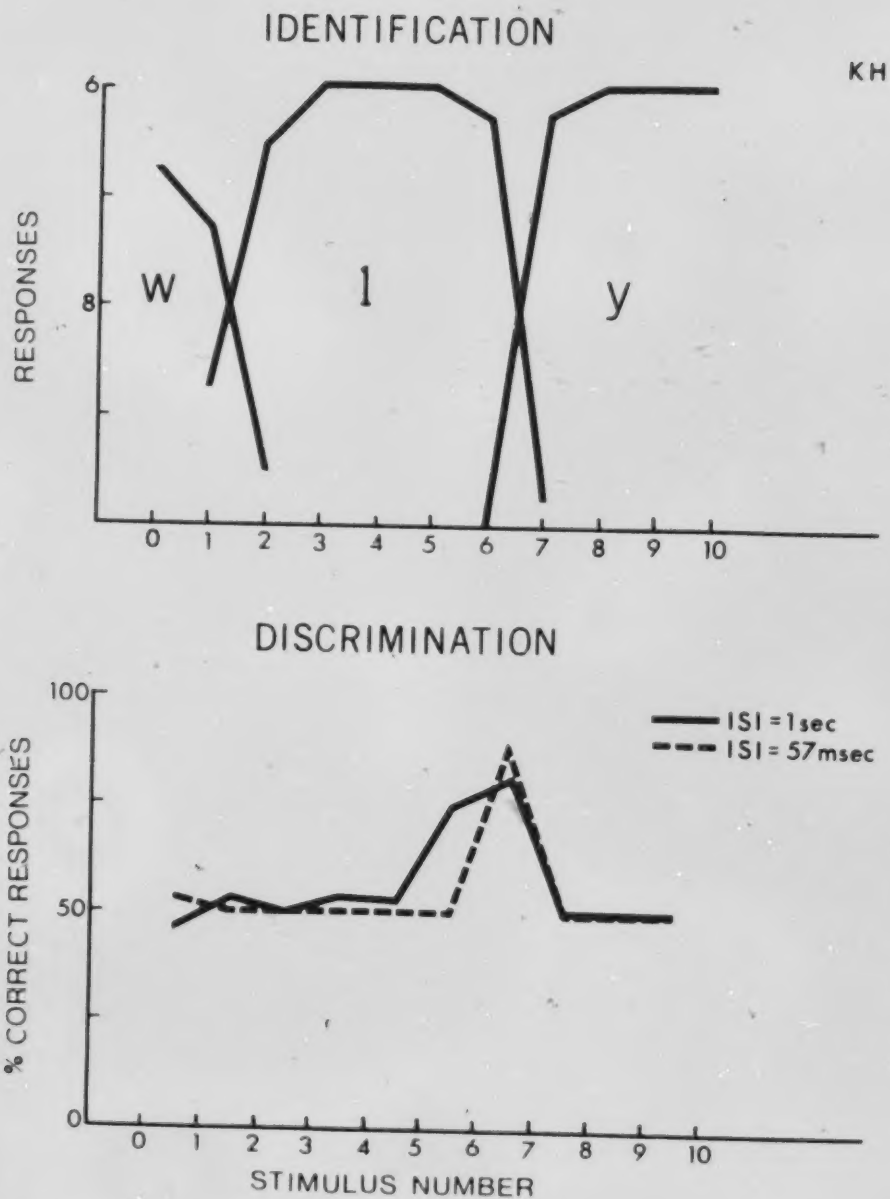


Figure 3: Identification and percent total correct discrimination responses for subject KH (Experiment III).

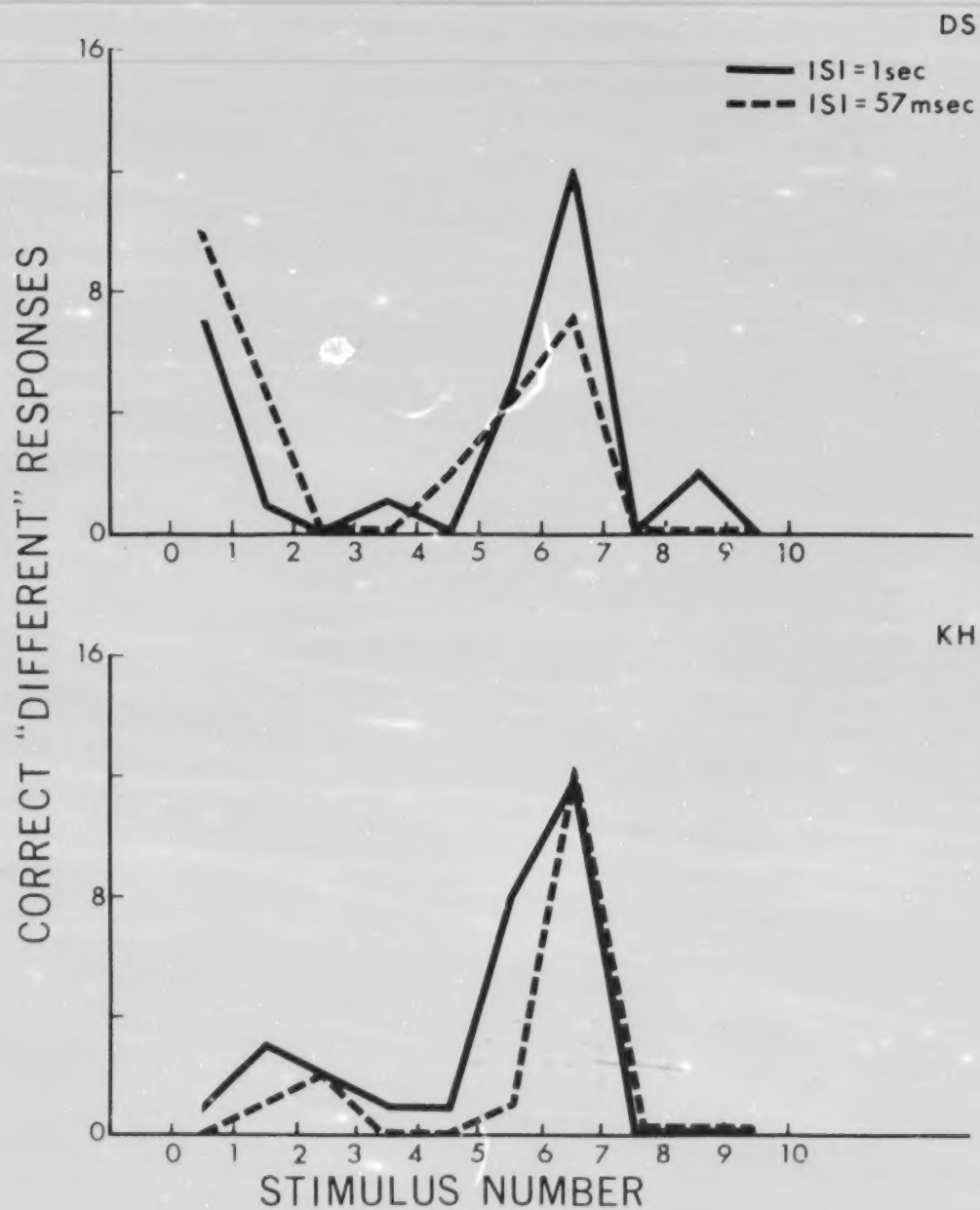


Figure 4: Number of correct different responses for subjects DS and KSH (Experiment III).

### Results of Experiments IV, V, VI, and VII

The mirror-image stimuli used in these experiments differ from the stimuli used in Experiments I, II, and III in that they do not all occur in American English. Stimulus /ɛw/ does not occur in any dialect of American English, whereas /ɛl/ occurs as an acceptable sequence in all dialects, and /ɛy/ is similar to the diphthongs used in at least some dialects. With this in mind, let us now consider the results of these experiments.

The data from all but one of the 24 subjects who participated in Experiments IV-VII showed that the perception of the mirror-image stimuli was also categorical, in the sense described in the last section. Figures 5 and 6 show the responses of two subjects who participated in Experiment VI.

Although the identification curves shown in Figures 5 and 6 are typical of most of the subjects who participated in Experiments IV-VII, not all subjects were this consistent in their identification of the stimuli, particularly at the lower values of  $F_2$ . As in Experiments I-III, the location of the /w/-/l/ category boundary showed more variation than the location of the /l/-/y/ category boundary.

### Comparison of the Results from Experiments I-VI

The group data for Experiments I-VI is shown in Table 2 in numerical form, because it was thought that this form would be the most useful for discussion (and for reference by subsequent investigators). The data from Experiment VII were not included in Table 2 because they were consistent with the other data, and because their inclusion would have prevented comparison of the two ISI conditions (because the subjects in Experiment VII took only the short ISI discrimination test).

Table 2 shows that the total number of correct responses for each of the experimental conditions is approximately the same. However, there are somewhat more correct "different" responses for the mirror-image stimuli in both ISI conditions. There are also more false "different" responses for the mirror-image stimuli.

### DISCUSSION

The results of Experiments I-VII reveal that the perception of /w/, /l/, /y/ is categorical in all experimental conditions. These results are consistent with Pisoni's (1973) results and with the Fujisaki and Kawashima explanation of categorical perception, because the number of correct discrimination responses within phonetic categories was not reduced by increasing the ISI to one-sec (which, as Pisoni demonstrated for vowels, is sufficient time for auditory information to decay). Thus these results support the hypothesis that auditory information is not available for speech sounds that are perceived categorically.

These results are in direct conflict with the "masking" hypothesis which predicted that the mirror image stimuli would be perceived continuously at a

Figure 5: Identification and percent total correct discrimination responses for subject KY (Experiment VI).

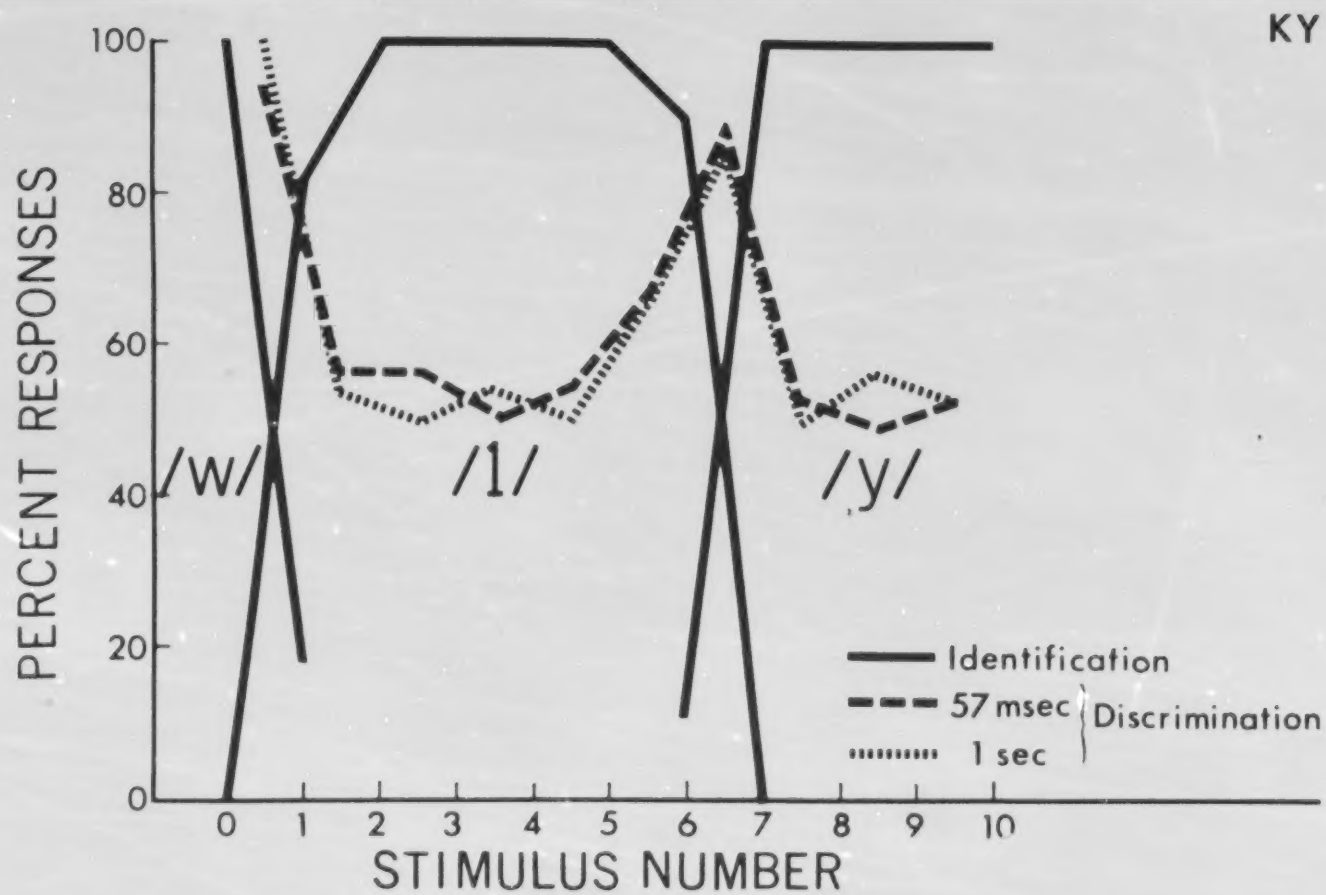




Figure 6: Identification and percent total correct discrimination responses for subject MK (Experiment VI).

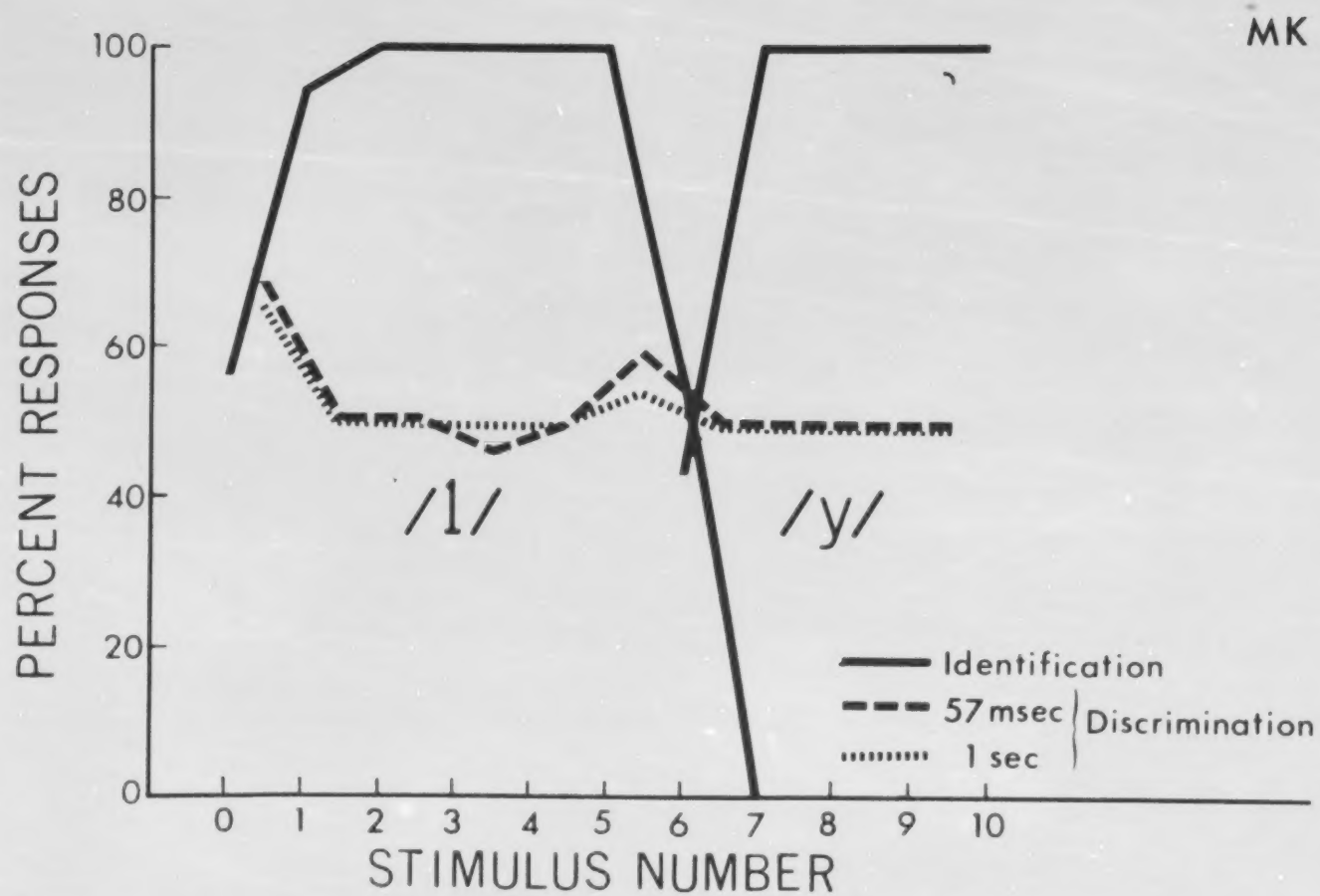


TABLE 2: Combined responses for Experiments I-VI.

|                      |  | Total Correct Responses                                                              |                         |                        |                          |
|----------------------|--|--------------------------------------------------------------------------------------|-------------------------|------------------------|--------------------------|
|                      |  | (each square represents 12 subjects x 32 responses = 384 possible correct responses) |                         |                        |                          |
| Stimuli:<br>Discrim: |  | /w/-/ye/<br>1 sec=ISI                                                                | /w/-/ye/<br>57 msec=ISI | /ew/-/ey/<br>1 sec=ISI | /ew/-/ey/<br>57 msec=ISI |
| Stimulus             |  |                                                                                      |                         |                        |                          |
| 0                    |  |                                                                                      |                         |                        |                          |
| 0<1                  |  | 222/384                                                                              | 240/384                 | 291/384                | 291/384                  |
| 1<2                  |  | 237                                                                                  | 228                     | 203                    | 197                      |
| 2<3                  |  | 193                                                                                  | 192                     | 201                    | 199                      |
| 3<4                  |  | 197                                                                                  | 193                     | 200                    | 199                      |
| 4<5                  |  | 210                                                                                  | 106                     | 201                    | 199                      |
| 5<6                  |  | 255                                                                                  | 223                     | 246                    | 237                      |
| 6<7                  |  | 259                                                                                  | 246                     | 282                    | 275                      |
| 7<8                  |  | 204                                                                                  | 199                     | 198                    | 203                      |
| 8<9                  |  | 201                                                                                  | 192                     | 190                    | 198                      |
| 9<10                 |  | 200                                                                                  | 192                     | 200                    | 195                      |
| Total                |  | 2178                                                                                 | 2101                    | 2212                   | 2193                     |

|          |  | Correct "Different" Responses                                                                    |        |         |         |
|----------|--|--------------------------------------------------------------------------------------------------|--------|---------|---------|
|          |  | (each square represents 12 subjects x 16 responses = 192 possible correct "different" responses) |        |         |         |
| Stimulus |  |                                                                                                  |        |         |         |
| 0-1      |  | 44/192                                                                                           | 53/192 | 106/192 | 106/192 |
| 1-2      |  | 61                                                                                               | 57     | 30      | 32      |
| 2-3      |  | 11                                                                                               | 34     | 12      | 25      |
| 3-4      |  | 7                                                                                                | 8      | 13      | 21      |
| 4-5      |  | 19                                                                                               | 6      | 17      | 17      |
| 5-6      |  | 70                                                                                               | 33     | 67      | 50      |
| 6-7      |  | 83                                                                                               | 72     | 112     | 85      |
| 7-8      |  | 19                                                                                               | 19     | 12      | 15      |
| 8-9      |  | 11                                                                                               | 7      | 7       | 9       |
| 9-10     |  | 12                                                                                               | 2      | 13      | 4       |
| Total    |  | 337                                                                                              | 291    | 389     | 364     |

|          |  | False "Different" Responses                                                                    |       |       |       |
|----------|--|------------------------------------------------------------------------------------------------|-------|-------|-------|
|          |  | (each square represents 12 subjects x 16 responses = 192 possible false "different" responses) |       |       |       |
| Stimulus |  |                                                                                                |       |       |       |
| 0        |  | 13/192                                                                                         | 5/192 | 7/192 | 7/192 |
| 1        |  | 16                                                                                             | 21    | 19    | 27    |
| 2        |  | 10                                                                                             | 34    | 14    | 17    |
| 3        |  | 2                                                                                              | 7     | 16    | 15    |
| 4        |  | 1                                                                                              | 2     | 19    | 9     |
| 5        |  | 5                                                                                              | 2     | 13    | 3     |
| 6        |  | 16                                                                                             | 18    | 22    | 12    |
| 7        |  | 7                                                                                              | 12    | 6     | 4     |
| 8        |  | 2                                                                                              | 7     | 9     | 3     |
| 9        |  | 2                                                                                              | 2     | 5     | 1     |
| 10       |  | 5                                                                                              | 5     | 3     | 3     |
| Total    |  | 79                                                                                             | 115   | 133   | 101   |

short ISI.

Because the acoustic information that distinguished each of the 11 test stimuli from the others was present both in the 90 msec initial  $F_2$  steady-state and the 60 msec  $F_2$  transition, these results are also in direct conflict with the hypothesis that transience of information determines that a sound will be perceived categorically. The stimuli /w/, /l/, /y/ were perceived categorically even though the acoustic cues for these sounds lasted 150 msec.

The results of Experiments I-VII also illustrate a problem with the scoring method traditionally used in categorical perception studies. The traditional scoring method combines a subject's correct "same" responses with that subject's correct "different" responses, allowing an unusually high number of false "different" responses to cancel out an unusually high number of correct "different" responses. This "cancellation" is problematic because it can camouflage a subject's discrimination peak, and also because it obscures the possibility that both effects (the high number of correct "different" responses and the high number of false "different" responses) have the same cause.

If these two effects do have the same cause--a comparison of different categorization labels (either for two occurrences of the same stimuli or for two different stimuli)--then both effects should occur where identification is least consistent, and both should be minimized where identification is most consistent. If this interpretation of false "different" responses is correct, the explanation of the relatively large number of false "different" responses for the mirror-image stimuli is that they were not identified as consistently as the original stimuli. As mentioned earlier, the mirror-image stimuli in fact were not identified as consistently as the original stimuli, particularly at the lower  $F_2$  frequencies. Another look at the false "different" responses in Table 2 confirms that the largest number of false "different" responses were always given for the set of stimuli (stimulus 0-4) drawn from the lower frequency end of the continuum. However, before this conclusion can be drawn with any certainty, it should be established that a comparable number of false "different" responses are not given by subjects asked to discriminate steady-state vowels presented with a relatively short ISI.

Looking at the correct "different" responses in Table 2, more correct "different" responses were given when /w/, /l/, /y/ occurred in syllable-final position than when they occurred in syllable-initial position. One might suggest that this reflects the presence of more auditory information for these stimuli than for the syllable-initial stimuli, except that the effect should then be seen only in the "short" ISI condition--whereas, in fact, it is seen in both ISI conditions. What is most mystifying about this effect, at least given the current assumptions about the type of information present in phonetic memory, is that this effect is largely due to enhanced discrimination at the category boundaries, not intraphonemically. Apparently an explanation of this effect must await further research, perhaps directed at the assumption that phonemic memory stores only discrete information.

## CONCLUSION

In summary, the experiments reported in this study demonstrate that the perception of synthetic /w/, /l/, /y/ is categorical in both syllable-initial and syllable-final position, in both 1-sec and 57-msec ISI conditions.

These results were interpreted: (1) to suggest that transience of information is not a crucial determinant of mode of perception of speech sounds, (2) to support the explanation of categorical perception proposed by Fujisaki and Kawashima, (3) to strongly falsify a "masking" explanation of categorical perception, and (4) to illustrate a problem with the scoring method traditionally used in studies of categorical perception.

## REFERENCES

- Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper & Row).
- Crowder, R. G. (1972) Visual and auditory memory. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge: MIT Press), pp. 251-275.
- Cutting, J. E. and B. S. Rosner. (in press) Categories and boundaries in speech and music. Percep. Psychophys.
- Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Lang. Speech 5, 171-189.
- Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of The Engineering Research Institute, 28, Faculty of Engineering. (Tokyo: University of Tokyo), pp. 67-73.
- Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanisms. Annual Report of the Engineering Research Institute, 29, Faculty of Engineering. (Tokyo: University of Tokyo), 207-214.
- Lane, H. L. (1965) the motor theory of speech perception: a critical review. Psychol. Rev. 72, 275-309.
- Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Am. 29, 117-123.
- Liberman, A. M., F. S. Cooper, D. S. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
- Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.
- Lisker, L. (1957) Minimal cues for separating /w,r,l,j/ in intravocalic position. Word 13, 257-267.
- Mattingly, I. G., A. M. Liberman, A. L. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, no. 2, 151-157.
- Miller, J. D., R. E. Pastore, G. C. Wier, W. J. Kelly, and R. J. Dooling. (1974) Discrimination and labeling of noise-buzz sequences with varying noise-lead times. J. Acoust. Soc. Am. 55, 390(A).
- O'Connor, J. D., L. J. Gerstman, A. M. Liberman, P. C. Delattre, and F. S.



- Cooper. (1957) Acoustic cues for the perception of initial /w,j,r,l/ in English. Word 13, 25-43.
- Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, no. 2, 253-260.
- Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants. J. Acoust. Soc. Am. 51, no. 4 (part 2).
- Stevens, K. N. (1968) On the relations between speech movements and speech perception. Z. Phon., Sprachwiss. u. Komm. Fschg. 213, 102-106.
- Studdert-Kennedy, M. (1976) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press), pp. 243-293.

Laterality and Localization: A Right Ear Advantage for Speech Heard on the Left\*

Christopher J. Darwin,<sup>†</sup> Peter Howell<sup>†</sup>, and Susan A. Brady<sup>†</sup>

ABSTRACT

The question of whether ear differences in the perception of speech are based on the ear of entry of a sound or its subjective location is addressed, using a paradigm which dissociates the two. If the first formant of a syllable is played to one ear, and the second and third formants to the other ear, the sound is localized to the ear getting the first formant, even though the discriminanda for a place-of-articulation judgment are only the second and third formants. A same-different reaction time (RT) paradigm showed that there was a reliable, but small, ear difference favoring the right ear when it received the second and third formants, indicating that ear of entry is a significant factor. However, there was an additional advantage for the right ear attributable to the apparent location of the sound. This latter factor was confined to the "different" judgments. A model of the functional organization of the afferent auditory system is offered to account for these and other results and for the role of attentional mechanisms in lateral asymmetries for speech discussed here.

---

\*To appear in the proceedings of the conference, Attention and Performance VII, held at Senanque, France, August 1976.

<sup>†</sup>Laboratory of Experimental Psychology and Centre for Research on Perception and Cognition, University of Sussex, England.

Acknowledgement: The computing facilities used in this experiment were made available through a grant from the Medical Research Council. The two junior authors were supported by a grant from the Science Research Council to C. J. Darwin. A pilot version of this experiment was conducted at Haskins Laboratories and at the University of Connecticut by Susan Dutch, whom we thank. We also thank Dr. A. E. Ades for some useful comments.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

## INTRODUCTION

The existence of auditory perceptual asymmetries in normal human subjects has substantially confirmed the clinically based hypothesis that man's two hemispheres are complementarily specialized for recognizing verbal and nonverbal sounds (Kimura, 1961a, 1961b, 1964; Milner, 1961; Luria, 1966). The results of these laterality experiments have, in addition, raised questions about the mechanism that allows cerebral asymmetries to be revealed as differences in performance between the ears. In particular, the following questions have been raised: Is the perceptual asymmetry effect an advantage for a particular ear or for a particular half of space (Morais and Bertelson, 1973, 1975; Morais, 1974, 1975)? Under what conditions is dichotic competition necessary to reveal the effect (Kimura, 1961a, 1961b; Milner, Taylor, and Sperry, 1968)? Does the effect arise because of a stronger contralateral afferent projection by the auditory system, or because of an attentional bias towards one side (Treisman and Geffen, 1968; Kinsbourne, 1970)? Answers to these questions would be interesting, not merely for the light that they might shed on the origin of the ear difference effect itself, but more generally, for their implications for the functional organization of the auditory perceptual system (see Haggard, in press). This paper uses a recently developed technique for revealing perceptual asymmetries monaurally (Morais and Darwin, 1974; Morais, 1975) to investigate primarily the first of these questions, but with results that have implications for all three.

Kimura's original papers (1961a, 1961b, 1967) on asymmetries in the recall of digit strings, emphasized the need for dichotic presentation and described a physiologically based model to explain the need for dichotic presentation. This proposed that the ear difference effect is due to the slight predominance of the contralateral over the ipsilateral pathway found for monaural stimuli being exaggerated by dichotic competition. Physiological support for this theory is the demonstrably greater cortical evoked potential, contralateral to the ear stimulated (Rosenzweig, 1951), and the paucity of single cortical units firing only to the ipsilateral ear (Hall and Goldstein, 1968). Although Kimura's theory explicitly referred to superior contralateral pathways between ears and hemispheres, electrophysiological evidence had also shown greater cortical activity contralateral to the ear receiving the earlier or the louder of two clicks presented, one to each ear (Rosenzweig, 1954; Brugge and Merzenich, 1973). This suggests that spatial location might be a more relevant (though less objective) dimension than ears. Using stereophonic and diotic presentation, Morais and Bertelson (1973, 1975) have succeeded in showing effects similar to Kimura's, where both ears get both stimulus channels. In their most striking example, they find that recall for the digit string subjectively located on the right is superior to that on the left when position is cued only by a difference in time of arrival between the ears. It is difficult to reconcile this result with a model that invokes a stronger contralesional projection between ears (as such) and hemispheres, but as Morais (1974, 1975) points out, it can be reconciled to Kimura's general model if one assumes that the projection to the hemispheres is based on a spatial location computed at a low level of the auditory system, rather than on actual ear of arrival (see also Haggard, in press).

What then is spatial location: the real position of the sound source with respect to the subject's head, or the apparent position of the sound source to the subject? In an ingenious experiment Morais (1974, 1975) showed that both these were important. He misled subjects into thinking that sound came from a pair of visible, but mute loudspeakers, when it in fact came from a similar pair of hidden loudspeakers that could be in a different place from the visible ones. Morais found that both the audible and the visible speakers had to be at right angles to the subject to obtain a significant ear difference in the recall of two channels of simultaneously presented syllables.

To explain this result Morais proposed a duplex theory, maintaining that the ear difference effect was determined both by a structural mechanism, such as Kimura's, that was sensitive to the actual position of the sound source, and also to biased attentional processes, such as those proposed by Treisman and Geffen (1968) and Kinsbourne (1970), that were sensitive to the apparent position of the source. Although this provides an adequate explanation of Morais' own experiment, other dichotic experiments are more difficult to handle with this duplex theory.

Halwes (1969) played subjects stop vowel syllables dichotically--either on the same pitch or on different pitches. The subjective impression of a dichotic pair of syllables with the same vowel and pitch is of a single midline source, whereas with different pitches it is of two separate sources localized towards their respective ear. Despite this impression of two sources, Halwes found no greater ear advantage when the pitch differed on the two ears. Here, then, subjective localization does not apparently influence the size of the ear difference. In a similar vein, Kirstein (1970) has found that although temporally misaligning syllables makes them easier to attend selectively, it decreases the ear difference. However, in both these experiments it might be argued that the treatment that made attentional processes easier was also reducing the dichotic competition necessary on Kimura's model in order to reveal ear differences, leading to no net change in the size of the effect. This objection applies perhaps less strongly to an experiment by Darwin (1969, experiment 8) who looked at the size of the ear difference for syllable-final stop consonants preceded by different lengths of vowel. The formant transitions cueing the stops were always temporally aligned, but because of the different vowel lengths, the onsets of the sounds were staggered. Subjects reported hearing the same sound in either ear (although they were always different) more often when the vowels were the same length, but this condition also showed the greatest probability of the reported sound being from the right ear. Here, then the conditions that favor a single percept also favor the larger ear difference.

In the experiment reported here, we examine the relative importance for the ear difference effect of the apparent position of sound source and of the actual ear at which the stimulus arrives. We do this by using a paradigm that dissociates the two, allowing subjects to hear on the left, say, a sound that actually came to the right ear. This is made possible thanks to two previous findings.



First, Rand (1974) has shown, following earlier findings on fusion (Broadbent and Ladefoged, 1957) that when the first formant of a stop consonant-vowel syllable is presented to one ear and the second and third formants to the other, not only do the two sounds fuse into a single subjective percept, but the amplitude of the second and third formants can be reduced by at least 20 dB before subjects cease to be able to identify the consonant. With this attenuation, the single percept is heard localized well to the side of the first formant. Since the place of articulation of stop consonants can be cued entirely by changes in the second and third formants, it is possible, using Rand's split-formant technique, to produce the impression of a syllable arriving at one ear, while allowing the information that determines its place of articulation to enter only the opposite ear. Second, Morais and Darwin (1974) found a reliable ear difference when subjects judged whether the initial consonant of a syllable presented monaurally was the same or different from that of an earlier syllable presented binaurally. Their "different" responses were on the average 15 msec faster when the second sound came to the right ear rather than to the left ear.

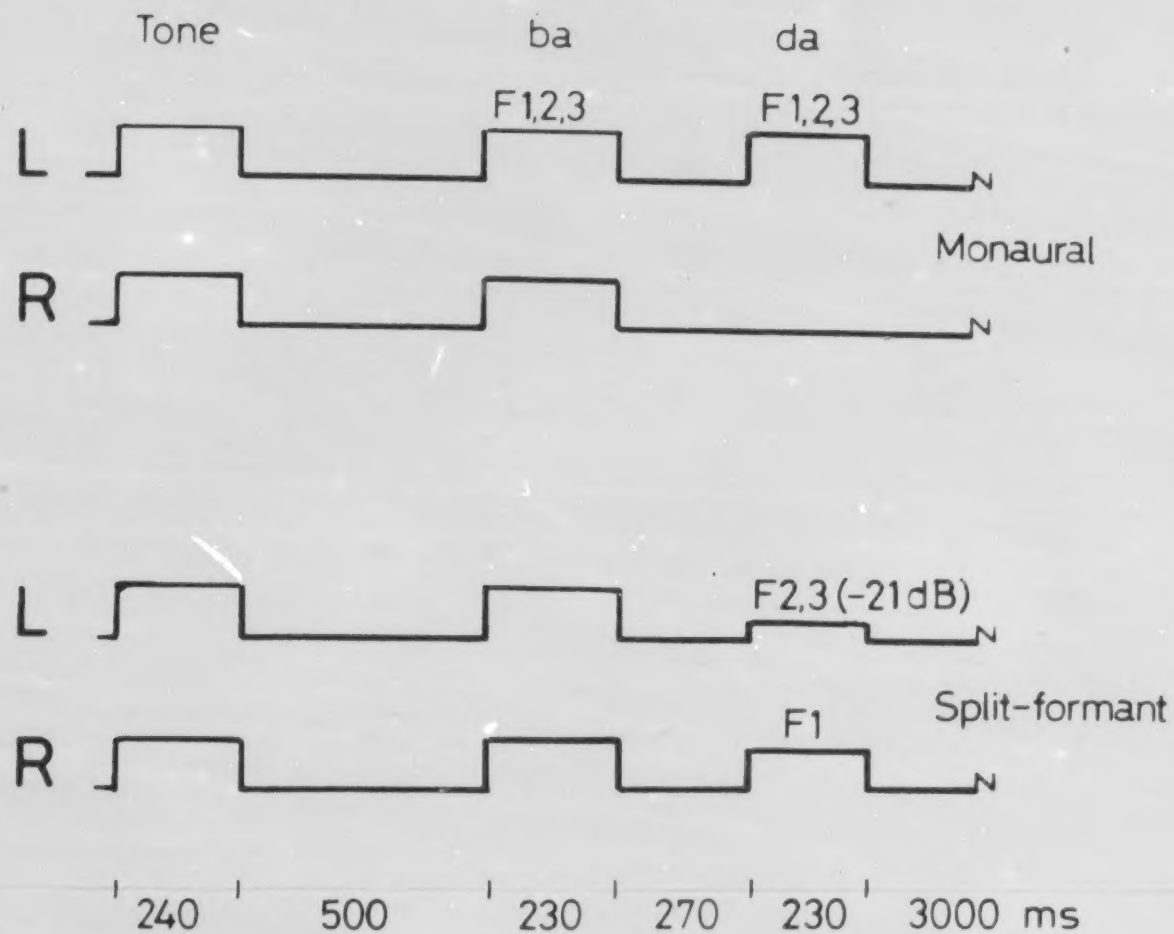
If, instead of playing the second syllable monaurally, we play a split-formant sound, will faster reaction times (RTs) appear in the case where the sound is heard on the right, or in the case where the discriminanda for the perceptual judgment come into the right ear? Kimura's original model predicts that actual ear of entry of the discriminanda should be important; Kinsbourne's attentional model and Morais' modification of Kimura's model predict that apparent location should be important, while Morais' duplex model predicts that both should be important.

#### METHOD

##### General

On each trial of the experiment the subject had to indicate as rapidly as possible whether two consecutively presented syllables had the same or a different initial consonant. He did this by moving a lever either away from or towards himself. Each trial started with a binaural warning tone and was followed by the first syllable presented binaurally on a low falling pitch. Half a second after the start of this syllable, the second one was played at a higher but also falling pitch. This second syllable could be played, either monaurally or with split-formants. In the monaural condition the whole syllable was presented to one ear, while in the split-formant condition the first formant was played to one ear simultaneously, with the second and third formants being played to the other ear 21 dB lower than they were in the monaural case. The combined level of the second and third formants was approximately 34 dB less than the first formant. The experimental format for a single trial in each of the two conditions is shown in Figure 1. There was a 3 sec gap between trials. Different pitch levels were used on the two syllables since this arrangement had been used in the earlier experiments of Morais and Darwin (1974).

Figure 1: Format of a single trial in the monaural and split-formant condition.



## Stimuli

Each syllable lasted 230 msec and was either /ba/, /da/, or /ga/, the consonant being cued by 40 msec transitions at the beginning of the syllable (Figure 2). For all three consonants the first formant rose from 350 Hz to a steady-state of 763 Hz, and the second and third formants had steady-state values of 1618 and 2545 Hz respectively. The starting frequency of the higher two formants, however, changed with the consonant and were (in the order /b/, /d/, /g/) 1373, 1753, 2100 Hz for the second formant, and 2200, 1750, 2200 Hz for the third formant. The binaural sounds' pitch fell from 121 to 92 Hz, while the monaural and split formant sounds' pitch fell from 151 to 122 Hz.

The synthesis of the speech sounds and the timing and output of the dichotic tape sequence were performed on an Elliott 4130 computer at the University of Sussex. The computed waveforms were stored on a disc, from which they could be retrieved and output simultaneously through a pair of D-A converters at a sampling rate of 7300 Hz. Each channel was lowpass filtered at 3 kHz (48 dB/oct) before recording on a Revox tape recorder. The stimuli were played back over Grason-Stadler TDH-39 headphones in a soundproof booth.

## Experimental Design

Each of 16 right-handed subjects, between ages 19 and 28, none of whom were known to have any hearing defect, attended five sessions, each lasting about two hours. The sessions were usually held on different days and no subject had more than two on the same day. Each session consisted of a warm-up block of 63 trials followed by four experimental blocks of 126 trials. The entire first session was a practice session whose data were discarded, leaving four scored sessions. Within each session, the four experimental blocks consisted of two blocks of the monaural condition and two blocks of the split-formant condition, with each condition using the trial order for its respective blocks. There were equal numbers of same and different trials, and the higher formants went to each replay channel equally often.

As well as the main experimental variable of monaural versus split-formant, three other variables were counterbalanced both within and between subjects: which hand held the response lever, headphone orientation, and the order in which the experimental blocks were heard in a session. In addition, the direction of movement to signal "same" was counterbalanced between subjects. So any individual subject always moved the lever the same way throughout the experiment to signal "same," but which hand he held it in varied from block to block.

At the start of a session, subjects took three identification tests, one binaural, one monaural, and one split-formant. Each test started with a demonstration of the three different syllables played twice in the order /ba/, /da/, /ga/, followed by a random sequence of 24 items which they had to identify. Two subjects were replaced at this stage of the experiment for failing to identify the sounds at better than 90 per cent.

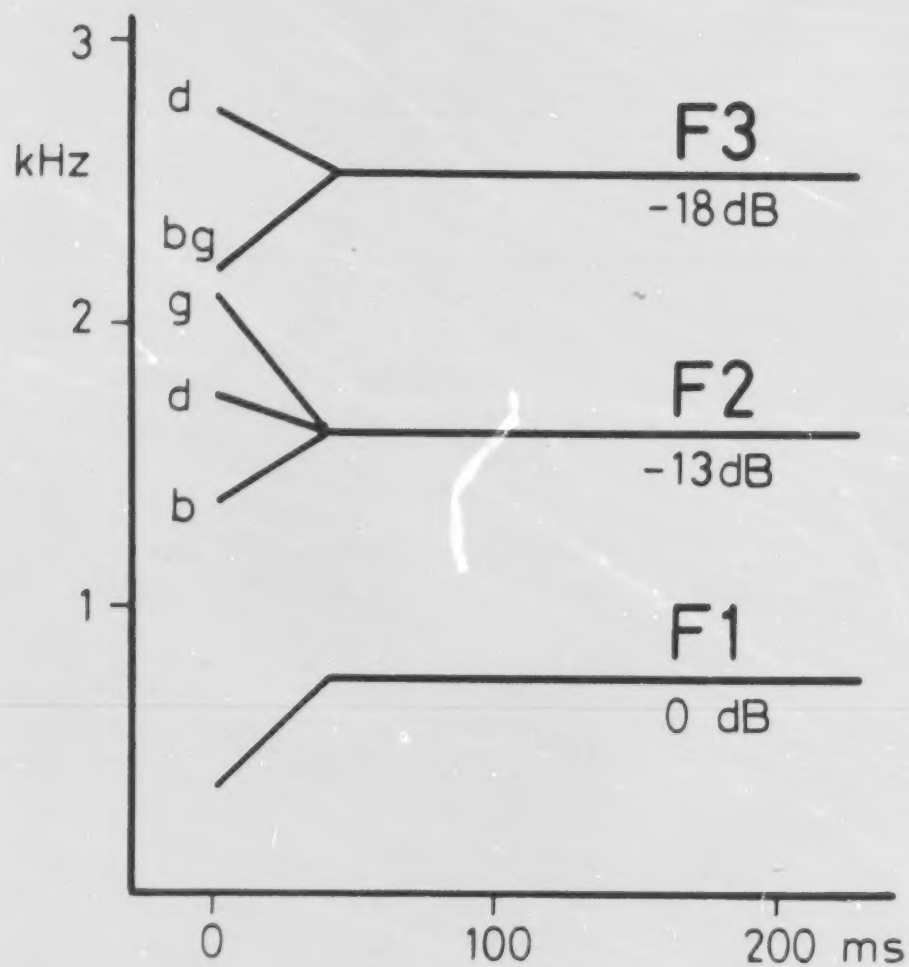


Figure 2: Stylized spectrograms for the three syllables /ba/, /da/, /ga/ used in the experiment. The amplitude levels are those used for the binaural and monaural stimuli. The higher formants were attenuated by an additional 21 dB in the split-formant condition.



The RT experiment was run with the help of a PDP-12 computer that started and stopped the tape recorder, recorded subjects' responses and reaction times and gave feedback to the subject via an oscilloscope display. The subject was told to indicate by a rapid lever movement whether the two syllables that he heard on each trial had the same or different initial consonants, and that he would get feedback via an oscilloscope screen which he could see through the window of the soundproof booth. During the practice block at the beginning of each session, the subjects saw "anticipation" whenever they made a response with a latency of less than 100 msec from the onset of the second syllable, or "wrong" whenever they made a mistake. During the experimental blocks (in both the practice and experimental sessions) they were additionally told whether their correct responses were "fast" or "slow," if a response was faster or slower than the mean of the preceding block. In addition to a flat rate of 70 pence per hour, subjects were paid 1/4 pence for each "fast" response and penalized 1/2 pence for each incorrect response; this gave them an average bonus of about 30 pence per session.

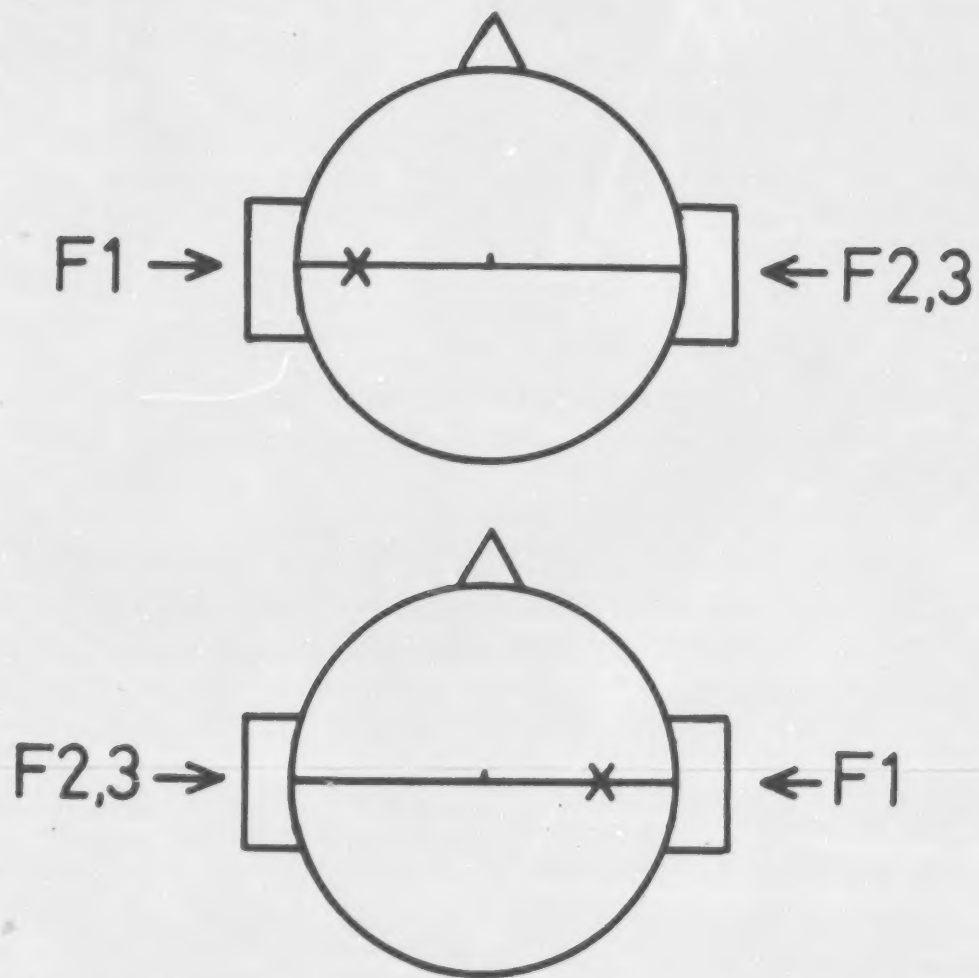
After the first and last experimental sessions subjects were asked to write down on which side they heard the split-formant sounds localized during a sequence of 63 trials, as a check that our own impression of these sounds' localization was shared by the subjects. Six of the 16 subjects also took a more elaborate localization test. They were given a diagrammatic head and asked to indicate on it the location of any sound they heard during the presentation of the second syllable. Each subject heard 60 split-formant trials. The instructions specifically emphasized that there might be more than one sound present.

## RESULTS

### 1. Identification and Localization Tests

The results of the identification tests confirmed Rand's (1974) finding that the split-formant sounds with extremely attenuated higher formants are readily identifiable. With binaural presentation, subjects averaged 98.5 percent correct, with monaural 98.8 percent, and with split-formant 97.6 percent. The 1.2 percent drop in intelligibility from monaural to split-formant condition was not significant across subjects. In the localization test taken by all subjects, every one reported hearing all the split-formant sounds localized to the side that was in fact getting only the first formant.

In the more elaborate test taken by six subjects, no one indicated more than one sound on any trial, and there were only two occasions when a subject marked a side other than that on which the first formant was presented. The mean apparent position of the sounds in the latter test are shown on the head in Figure 3. Although dichotic presentation of sounds with the same periodicity but grossly different spectral ranges can give rise to double images, this was not the case in our experiment. We have noticed, however, that if the amplitude of the second and third formants is raised above the considerably attenuated level used here, a double image can be heard (cf.



## Subjective Locale

Figure 3: Subjective location of the split-formant sounds for the six subjects who took the more elaborate localization test.

Cutting, 1976, experiment V).

## 2. Reaction Times

The raw data for the analysis were the mean reaction times for correct trials on each experimental block (of 120 scored trials). Wrong responses ("same" for "different" and the reverse) were counted as errors and, together with anticipations, were not included in the RT analysis. Correct responses with a latency of longer than 1 sec were also excluded from the RT analysis, although they were not counted as errors.

The mean RT's for the main experimental variables are shown in Figure 4. The basic result is that in both the monaural and the split-formant conditions there is an overall advantage for the right ear when this receives the second and third formants. With split-formants the faster condition is when the sound is localized to the left, whereas monaurally the faster condition is when the sound is localized to the right. The main statistical analysis was a six-way analysis of variance with factors: (1) experimental condition (monaural versus split-formant), (2) ear receiving second and third formants (left/right), (3) response (same/different), (4) direction of lever movement (nested under subjects), (5) replications (four sessions) and (6) hand (left/right). Significant main effects were attributable to the right ear being faster than the left ( $F_{1,14} = 17.3, p < .001$ ), same responses being faster than different ( $F_{1,14} = 19.1, p < .001$ ), monaural being faster than split-formant ( $F_{1,14} = 16.9, p < .005$ ) and a general speeding up over successive sessions ( $F_{3,42} = 9.1, p < .001$ ). No interaction reached significance at the 2.5 percent level except that between experimental condition, ears, and response ( $F_{1,14} = 24.1, p < .0005$ ), that is, between the dimensions given in Figure 4, and more clearly in Figure 5. This interaction can be described as due to the following difference between the monaural and the split-formant conditions: in the monaural condition the greater ear difference is seen for the different responses, whereas in the split-formant condition the greater ear difference is for same responses. Separate analyses of variance on the monaural and split-formant conditions confirmed this conclusion. These showed a significant interaction between ears and response, both for monaural ( $F_{1,15} = 12.6, p < .005$ ) and split-formant ( $F_{1,15} = 8.0, p < .025$ ) conditions, and also for both conditions, that same judgments were faster than different at the 1 percent level. Four analyses on the scores on either ear for same and different responses under either monaural or split-formant condition showed separately that in each of these conditions there was a significant advantage for the right ear:

monaural-same:  $F_{1,15} = 8.3, p < .025$

monaural-different:  $F_{1,15} = 26.0, p < .001$

split-formant-same:  $F_{1,15} = 19.2, p < .001$

split-formant-different:  $F_{1,15} = 5.4, p < .05$ .

In summary, the RT data show that subjects judge the second of two syllables to be the same or different from the first one faster when those aspects of the stimulus on which this judgment is based entered the right

Figure 4: Mean reaction times (RT) for same and for different judgments made in the monaural and the split-formant conditions, displayed according to the ear receiving the second and third formants.

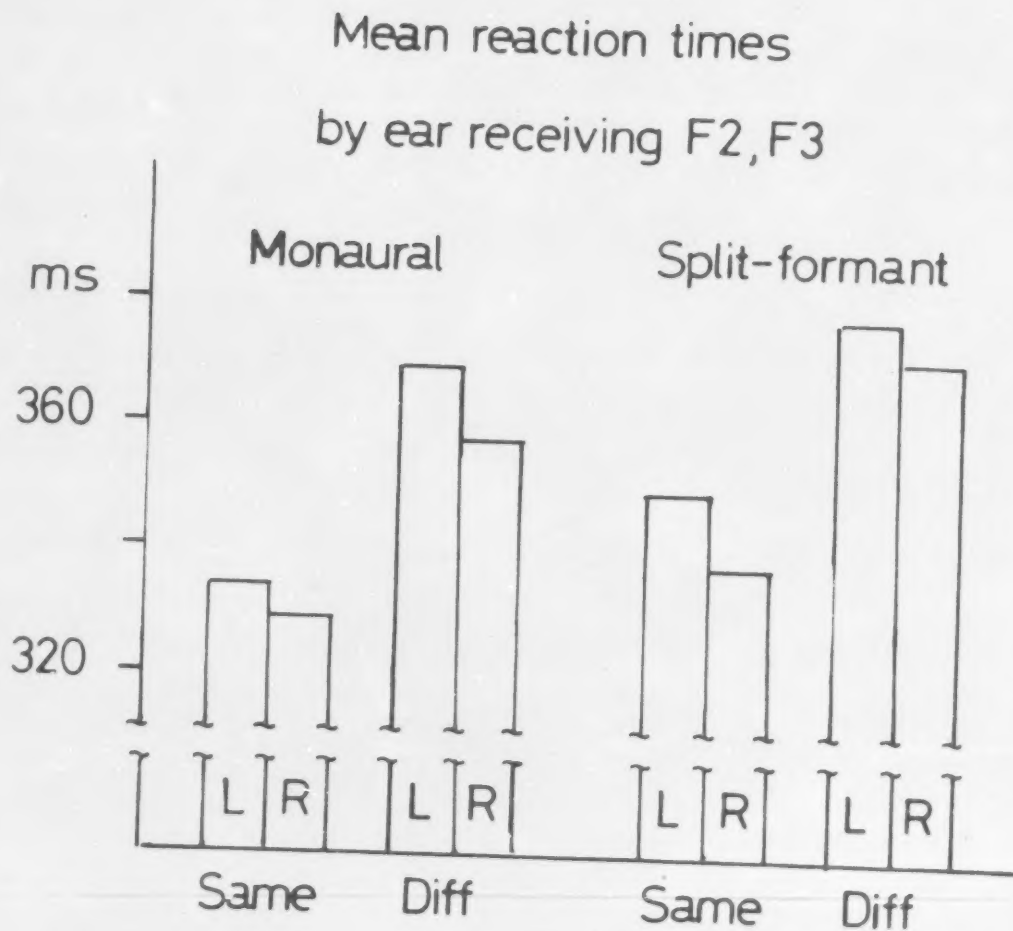
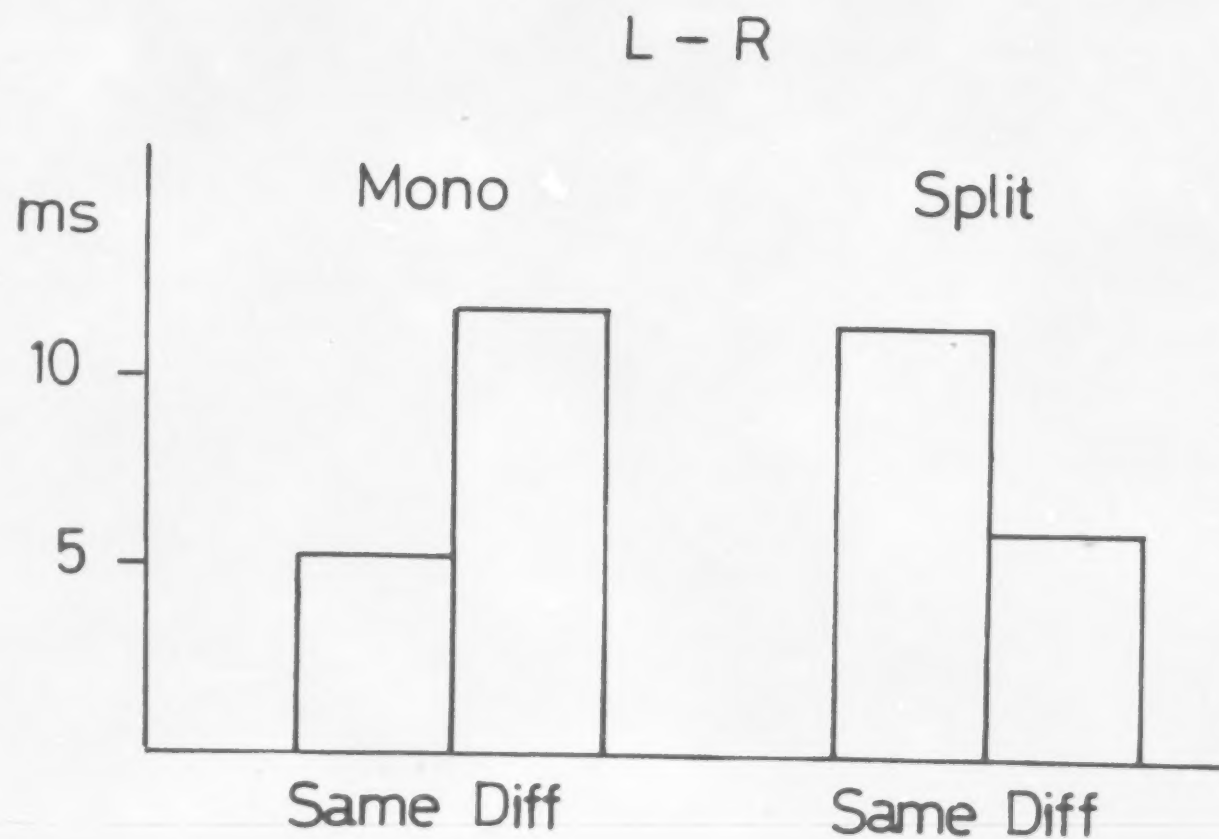




Figure 5: Mean difference in reaction-time (RT) between the ears for same and different responses in the monaural and split-formant conditions.



ear, than when they entered the left. This was true whether subjects heard the sound as being on the left or on the right. The size of this ear advantage was monaurally greater for different than for same responses (replicating Morais and Darwin, 1974), but the opposite was true in the split-formant case. In general, same responses were faster than different responses (and this did not change with splitting the formants), and RTs were faster in the monaural condition than in the split-formant.

### 3. Errors

The number of responses that were errors (that is, wrong responses and anticipations) was also calculated for each block, and the means for the three main experimental variables are shown in Table 1. An analysis of variance was performed on this data after subjecting it to an arc-sine root transformation. It is important to see whether there are any ear differences in the error data, since if there are, and they are in the opposite direction to the RT data, then the interpretation of both error and RT data is complicated by the fact that accuracy and speed can be traded off against each other (Swensson and Edwards, 1971).

An analysis of variance showed that there was no main effect involving ear, nor any two-way interaction, but that there was a three-way interaction between ear, experimental condition, and hand ( $F_{1,14} = 7.0, p < .025$ ). The interaction can be described as follows: in the split-formant condition more mistakes are made by a given hand when the higher formants come into the ear on the same side as the hand, whereas for the monaural condition the reverse is true. Put another way, this indicates that for both conditions more mistakes are made by a particular hand when the sound is localized to the opposite side than when it is on the same side. Thus errors seem to be determined more by the sound's apparent location relative to the responding hand--in contrast to the main RT effect that is determined by the ear of entry of the higher formants irrespective of hand. Higher order interactions involving these same factors, however, restrict this effect to those subjects who made away movements for same responses, and within them to the same response. Errors then behave quite differently from the mean RT's, as they give much less general effects involving ear, which turn out to be based on a sound's apparent location rather than ear per se.

The error analysis also revealed two main effects and a two-way interaction not involving ear: those subjects who moved the lever away to respond "same" were more accurate than the other subjects ( $F_{1,14} = 12.1, p < .005$ ), though this effect interacted with whether the correct response should have been same or different ( $F_{1,14} = 6.3, p < .025$ ). In addition, as one might have expected from the identification and the RT data, slightly fewer errors were made in the monaural than in the split-formant condition ( $F_{1,14} = 8.5, p < .025$ ).

## DISCUSSION

This experiment has shown that when subjects have to move a lever to indicate whether the second of two syllables is the same as or different from the first, they make this movement sooner when the discriminanda distinguishing the syllables come into the right ear than if they come into the left. This is true regardless of whether the fused, complete syllable is heard on the left or the right side. In conjunction with the findings of Morais and Bertelson (1975), this poses problems for the three simple models of the origin of the ear-difference effect described in the introduction.

The two models which attribute the effect to a contralateral afferent projection either by ears or by spatial location, can each handle only one of the results. The ear advantage cannot be simply an ear advantage, because Morais and Bertelson (1975) found that when both messages were channelled to both ears, but with different temporal offsets to produce an impression of different localizations, the sound heard on the right was recalled better than that on the left. Nor can the ear advantage be simply a matter of subjective localization, since our experiment shows that when a single sound is split into high and low frequency components, the ear advantage is to the side receiving the high frequency sounds, while the localization is to the side receiving the low frequency sounds. A model that attributes the effect to a direction of attention to a contralateral half of space has similar problems. Where is being attended? If the contralateral half of subjective space is attended to, then the wrong prediction is made for our experiment; if ears are being attended, then the wrong prediction is made for the Morais and Bertelson experiment. How do we resolve the apparent contradiction between these two experiments?

Perhaps the answer lies in considering more carefully the sorts of fusion processes that must be taking place in Morais and Bertelson's stereophonic experiment and in our split-formant experiment. In their experiment, the two ears receive identical waveforms with a slight difference in relative time of arrival. This imitates, in part, the natural situation in which sound from a source to one side of the head reaches one ear before the other. In the natural situation, the higher frequencies will be attenuated somewhat at the ear away from the source, because of the acoustic shadow cast by the head. However, in general, all frequencies arriving at one ear from a particular source will be present, albeit at different times and amplitudes, at the other ear. In our split-formant experiment, though, there is virtually no spectral overlap between the two ears. It is possible that different mechanisms are operating to produce "fusion" in these two cases.

Binaural fusion could be achieved for sounds composed of the same spectral frequencies by a crosscorrelational process that only compared signals within the same narrow frequency band (Colburn, 1973; Toole and Sayers, 1965). Such a mechanism cannot, however, be the whole story, since sounds with a similar waveform envelope but virtually no overlap in their spectra will still fuse into a single percept (Broadbent and Ladefoged, 1957). Fusion by a common envelope cannot in turn be the prime mechanism,

since the fusion produced when there is no spectral overlap is less robust than that found when both the envelope and the spectrum are similar (Broadbent and Ladefoged, 1957; Cutting, 1976, exp. V; Toole and Sayers, 1965). In addition, Henning (1974) has found that the ability to localize a fused image on the basis of time differences between the envelopes of different sounds at the two ears deteriorates rapidly as the spectral similarity between these signals is reduced.

It is possible, though not compelling, that there are two distinct mechanisms responsible for fusion, one that looks only at corresponding spectral frequencies, and one that is concerned with envelope similarities (cf. Sayers and Cherry, 1957; Toole and Sayers, 1965). This distinction has been made in the context of laterality experiments by Haggard (in press), who speculates that "fusion and assignment to location between and within hemispheres on a spatial basis takes place at a lower level in terms of spectral similarities than it does in terms of shared periodicities" (Haggard, 1975). We will now elaborate this notion and show that it can account for the discrepancy between our experiment and Morais and Bertelson's.

First, let us see how this distinction between different levels of fusion can be applied to Kimura's model and Morais' modification of it. Both these models require some contralateral superiority of the afferent auditory projection; they differ in whether contralateral refers to ears or to spatial location. Let us assume that Morais is correct when he suggests that the contralateral projection is based on the result of some comparison process of the signals at the two ears, rather than being based on each ear alone. Let us now assume, following Haggard (1975), that this comparison process only involves a combination of sounds within the same frequency band. On the basis of such a process each spectral component of a signal would be projected predominantly to higher anatomical levels that are contralateral to a position calculated by comparing the two ears for that frequency band. Split-formant sounds would not be fused at this level but would be treated, despite their identical pitches, as two distinct sounds each projecting predominantly to the side contralateral to their initially calculated position (that is, contralateral to the side of their respective ear of arrival). Morais and Bertelson's sounds on the other hand, would, at this level, be associated with the side at which they were then subjectively localized.

A similar resolution can be achieved for attentional models such as Kinsbourne's by assuming that attention is directed to the initial spatial location, rather than to the apparent subjective location of the sound.

#### Same-Different Effects

Splitting the formants of the second syllable in this experiment produced a highly significant change in the interaction between ears and in whether the subject responded "same" or "different." In the monaural condition, as in Morais and Darwin (1974), there is a significantly larger ear difference for different than there is for same responses. The present experiment collected more data and the small ear difference for same



responses reached significance. For the split-formant sounds, on the other hand, there was a significantly larger ear difference for same judgments than for different, and again both reached significance. The explanation originally offered for the significant ear difference for different judgments by Morais and Darwin, took advantage of a model proposed by Bamber (1969) to explain why, for readily categorizable stimuli (Bindra, Donderi, and Nishisato, 1968), same judgments are faster than different. This model maintains that while same judgments can be made solely on the basis of a low level representation of the stimulus, resort must be made to a higher categorical process to determine that stimuli are categorically different.

The need for this categorical processing yields longer RTs. Morais and Darwin explained their results by assuming that the categorical process involved in different judgments but not required by the same judgments, is represented more in the left than the right hemisphere, whereas same judgments can be made on the basis of an unlateralized mechanism. Although at first sight, the results of the present experiment appear to be contrary to this model, it does in fact provide the skeleton on which to build an explanation of our results. The important additional postulate is that there are two lateralized mechanisms. The first is responsible for making same judgments and does this faster for sounds which enter the right than the left ear. The second is required, in addition to the first, for different judgments and is faster for sounds subjectively localized on the right side.

Thus, for the monaural condition we have a small right ear advantage (REA) for same judgments, supplemented by an additional advantage for different judgments based on apparent location. In the split-formant condition, though, the additional ear difference for the different judgments favors the condition where the higher formants come to the left ear. Thus the REA is smaller than for same judgments. We have also to explain why the same judgments in the split-formant case show more of a REA than in the control case. A related observation is that RTs on the whole are slower in the split-formant condition than in the monaural condition. We might put forward the hypothesis then that comparing a split-formant sound with a binaural one requires another process in addition to simply comparing a binaural with a monaural, and that this process is sensitive to the location, determined on a spectral basis, of the component of the sound relevant to the discrimination. This process, which combines sounds on similar pitches at the two ears, but does not assign them to a new location, is defined in an identical way to Cutting's (1976) "spectral fusion."

Three separate factors need to be taken into account:

- (1) A process that is faster at handling sounds presented to the right side and that can judge sounds with identical formant structure to be the same.

- (2) A process that can judge sounds fused according to a common periodicity to be the same as binaurally presented sounds, and that also handles sounds faster when the component relevant to the discrimination is presented to the right side.

(3) A process that can judge speech categories to be different and one that handles sounds faster when they are subjectively localized to the right side.

Figure 6 illustrates how these processes might be linked together functionally.

Of the three lateralized processes used here to explain our results, only one--the process for categorizing speech sounds--needs to be expressly linguistic. Retaining and comparing sounds fused either by spectral similarity or common periodicity is not something uniquely useful in verbal classification. Why then should we find that it elicits an advantage for the right ear? This may be a further example of a context effect, similar to the shift in the ear advantage for vowels that can be produced by using all verbal material compared with a mixture of verbal and nonverbal material on different trials (Spellacy and Blumstein, 1970). A general explanation of such context effects can be made using Kinsbourne's (1970) notion of hemispheric arousal, but in our case it would be perhaps more parsimonious to assume that since the speech categorization process is being performed primarily in the left hemisphere, the results of other, bilateral processes are taken from the same hemisphere for a joint decision, because of better within (than between) hemisphere connections.

#### Dichotic Competition and Attention

This model has been developed mainly with the results of our split-formant and monaural RT experiment in mind. How does the model cater for the results of dichotic experiments, and in particular, what is the relative importance of ear of entry and subjective localization under dichotic competition?

With dichotic competition, reliable ear differences can easily be demonstrated in recall errors when subjects only report hearing a single sound source. In our monaural and split-formant experiments (see also Haggard, 1975; Perl and Haggard, 1974), by contrast, there is little indication of any ear difference in the error scores. Thus, the ear difference in recall errors due to the actual, rather than the apparent, sound location is what is enhanced by dichotic competition. Let us suppose, then, that the main effect of dichotic competition is selectively to attenuate the weaker pathway in Figure 6 leading from the initial spatial representation based on fusion within frequency bands. Errors under dichotic presentation will then depend almost exclusively on a sound's location in this initial spatial representation, rather than on its apparent position. Since our split-formant presentation is clearly not equivalent to dichotic presentation, we might also suppose that dichotic competition will only be effective for those frequency regions that are represented on either ear. Simply sharing frequency regions is not a sufficient condition, though, since white noise, although an effective competitor for clicks (Murphy and Venables, 1970), is not for speech (Coris, 1967; Darwin, 1971).

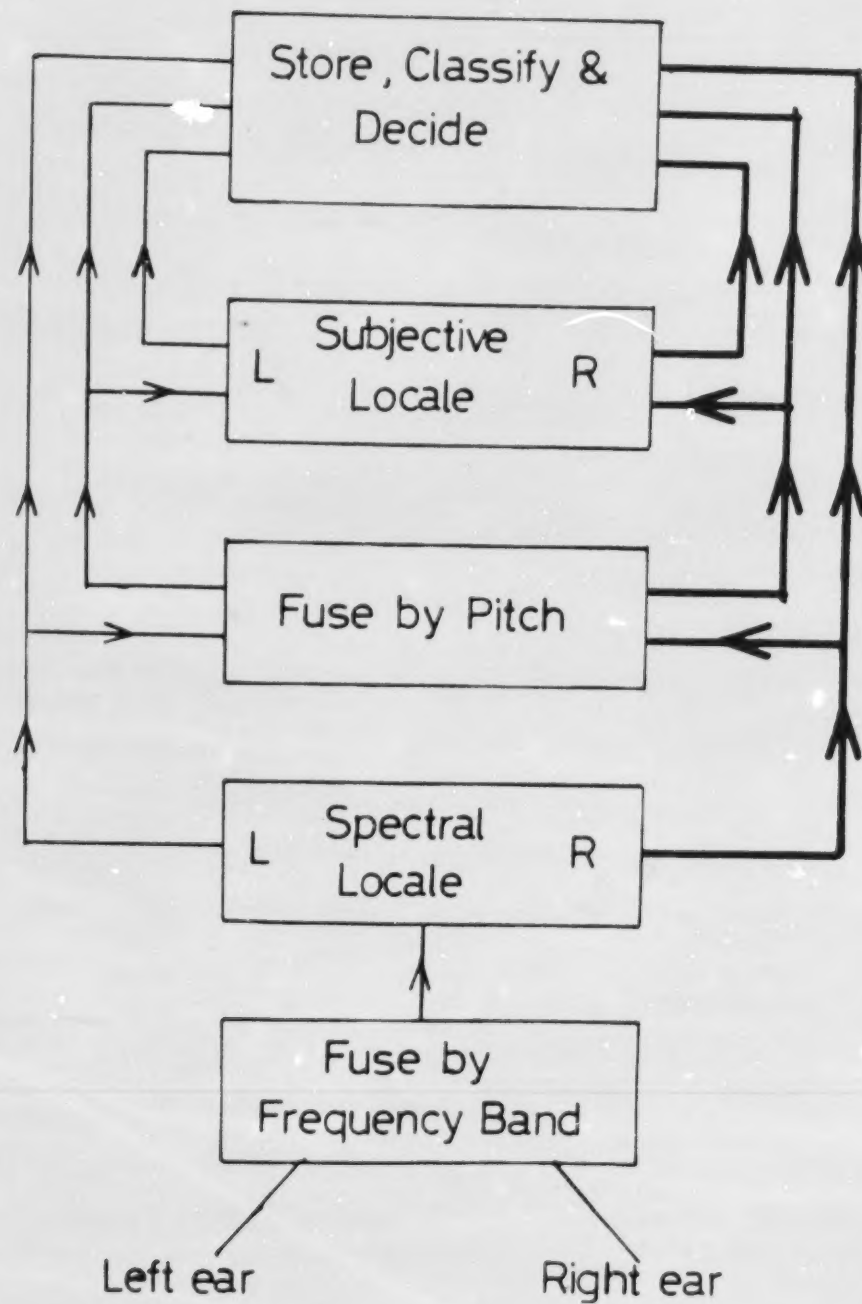


Figure 6: Model to illustrate how different levels of fusion could interact to give the ear differences observed in this experiment and in those using stereophonic presentation. "Same" judgments can be made on the basis of the output from the lower three boxes, while "different" judgments require, in addition, output from the box labeled Subjective Locale.

Since commissurotomed patients show a greatly reduced score on their left ear for recall of dichotically presented speech, but not for recall of monaurally presented speech (Milner, Taylor and Sperry, 1968; Sparks and Geschwind, 1968), the original Kimura hypothesis of ipsilateral suppression has been supported. Here we are speculating that it is the ipsilateral pathway from the initial spatial representation which is primarily inhibited by dichotic stimulation. The remaining uninhibited pathways would thus reach the ipsilateral hemisphere via the contralateral pathway and the corpus callosum. This organization would thereby imply that fusion by pitch and the determination of apparent location, were mainly cortical processes.

This model, as it stands, cannot handle Morais' (1974, 1975) finding that with loudspeaker presentation, the half-space difference for recall of strings of syllables was as much influenced by the real position of the sound source as by its imaginary position. As we have already remarked, this contrasts with Halwes' experiment using sounds on different pitches where the ear difference was found to be no greater under conditions favoring the identification of two sound sources. Morais' experiment differed from Halwes' both in using lists of syllables with different vowels--rather than a single pair of syllables with the same vowel--and real speech rather than synthetic. Morais' sounds thus differed in pitch and vowel quality at the ears as well as being extended in time. Although this should not result in any greater degree of subjective spatial separation than Halwes' different pitched sounds, it would provide additional dimensions for allowing the subject selectively to attend to one sound rather than the other. We might then agree with Morais (1974, 1975) that the influence of subjective location which he found under stereophonic competition is indeed the result of an attentional bias, but that this bias can only operate for sounds that are readily selected by being auditorially distinct and extended in time.

This necessary alteration to our original model raises the question of whether the effect of subjective location that we found in our experiment is best attributed to an attentional process, or to a more structural difference in the auditory pathways. There is at present no data that allows us to distinguish between these two alternatives.

In summary, then, the model that we propose retains aspects of all three of the models described in the introduction. The ear difference is in part determined by the ear of entry (or perhaps more precisely by the imaginary location of sound sources based on a comparison across ears within individual spectral regions), and in part by a sound's apparent location. This is true for our RT experiment. Recall errors in dichotic experiments appear to be mainly controlled by the ear of entry, but may perhaps be influenced by attentional mechanisms based on apparent location, if the sounds used allow easily selective attention. We tentatively suggest that the determination of a sound's apparent location is a cortical process and that the effect of dichotic competition is to inhibit only those ipsilateral pathways whose subsequent projections are based on a sound's ear of entry or spectrally determined spatial location.



# REFERENCES

- Bamber, D. (1969) Reaction times and error rates for "same"- "different" judgments of multidimensional stimuli. Percept. Psychophys. 6, 169-174.
- Bindra, D., D. C. Donderi, and S. Nishisato. (1968) Decision latencies of "same" and "different" judgments. Percept. Psychophys. 3, 121-130.
- Broadbent, D. E. and P. Ladefoged. (1957) On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Am. 29, 708-710.
- Brugge, J. F. and M. M. Merzenich. (1973) Responses of neurons in auditory cortex of the macaque monkey to monaural and binaural stimuli. J. Neurophysiol. 36, 1138-1158.
- Colburn, H. S. (1973) Theory of binaural interaction based on auditory-nerve data I, General strategy and preliminary results on interaural discrimination. J. Acoust. Soc. Am. 54, 1458-1470.
- Coris, P. M. (1967) The effects of contralateral noise upon the perception and immediate recall of monaurally-presented verbal material. Unpublished M. A. thesis, McGill University, Montreal.
- Cutting, J. E. (1976) Auditory and linguistic processes in speech perception: inferences from six fusions in dichotic listening. Psycholog. Rev. 83, 114-140.
- Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Ph. D. thesis, Cambridge University.
- Darwin, C. J. (1971) Dichotic forward and backward masking of speech and non-speech sounds. J. Acoust. Soc. Am. 50, 129(A).
- Haggard, M. P. (1975) Asymmetrical analysis of stimuli with dichotically split-formant information. Speech Perception Series 2, no. 4, Department of Psychology. (Belfast: The Queen's University of Belfast).
- Haggard, M. P. (in press) Dichotic listening. In Handbook of Sensory Physiology, ed. by R. Held, D. Leibowitz, and H. L. Teuber. (New York: Springer-Verlag).
- Hall, J. L. and M. H. Goldstein. (1968) Representations of binaural stimuli by single units in primary auditory cortex of unanaesthetized cats. J. Acoust. Soc. Am. 43, 456-461.
- Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Supplement to Haskins Laboratories Status Report on Speech Research, September.
- Henning, G. B. (1974) Detectability of interaural delay in high-frequency complex waveforms. J. Acoust. Soc. Am. 55, 84-90.
- Kimura, D. (1961a) Some effects of temporal lobe damage on auditory perception. Canadian J. Psych. 15, 156-165.
- Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canadian J. Psych. 15, 166-171.
- Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exper. Psych. 14, 355-358.
- Kimura, D. (1967) Functional asymmetries of the brain in dichotic listening. Cortex 3, 163-178.
- Kinsbourne, M. (1970) The cerebral basis of lateral asymmetries in attention. In Attention and Performance III, ed. by A. F. Sanders. (Amsterdam: North-Holland), pp. 193-201.

- Kirstein, E. (1970). Selective listening for temporally staggered dichotic CV syllables. Haskins Laboratories Status Report on Speech Research, SR-21/22, 63-70.
- Luria, A. R. (1966) Higher Cortical Functions in Man. (London: Tavistock), pp.
- Milner, B. (1961) Laterality effects in audition. In Interhemispheric Relations and Cerebral Dominance, ed. by V. C. Mountcastle. (Baltimore: Johns Hopkins), pp. 177-195.
- Milner, B., L. Taylor, and R. W. Sperry. (1968) Lateralized suppression of dichotically presented digits after commissural section in man. Science 161, 184-186.
- Morais, J. (1974-1975) The effects of ventriloquism on the right-side advantage for verbal material. Cognition 3, 127-139.
- Morais, J. (1975) Monaural ear differences for same-different reaction times to speech with prior knowledge of ear stimulated. Perceptual and Motor Skills 41, 829-830.
- Morais, J. and P. Bertelson. (1973) Laterality effects in diotic listening. Perception 2, 107-111.
- Morais, J. and P. Bertelson. (1975) Spatial position versus ear of entry as determinant of the auditory laterality effects: a stereophonic test. J. Exper. Psych.: Human Percept. Perform. 1, 253-262.
- Morais, J. and C. J. Darwin. (1974) Ear differences for same-different reactions to monaurally presented speech. Brain Lang. 1, 383-390.
- Murphy, E. H. and P. H. Venables. (1970) Ear asymmetry in the threshold of fusion of two clicks: a signal detection analysis. Quart. J. Exper. Psych. 22, 288-300.
- Perl, N. and M. P. Haggard. (1974) Masking versus hemisphere sharing of processing for speech sounds. Speech Perception Series 2, no. 3. Department of Psychology, (Belfast: The Queen's University of Belfast).
- Rand, T. C. (1974) Dichotic release from masking for speech. J. Acoust. Soc. Am. 55, 678-680.
- Rosenzweig, M. R. (1951) Representations of the two ears at the auditory cortex. Am. J. Physiol. 167, 147-158.
- Rosenzweig, M. R. (1954) Cortical correlates of auditory localization and of related perceptual phenomena. J. Compar. Physiol. Psych. 47, 269-276.
- Sayers, B. McA. and E. C. Cherry. (1957) Mechanisms of binaural fusion in the hearing of speech. J. Acoust. Soc. Am. 29, 973-987.
- Sparks, R. and N. Geschwind. (1968) Dichotic listening in man after section of neocortical commissures. Cortex 4, 3-16.
- Spellacy, F. and S. Blumstein. (1970) The influence of language set on ear preference in phoneme recognition. Cortex 6, 430-439.
- Swenson, R. G. and W. Edwards. (1971) Response strategies in a two-choice reaction task with continuous cost for time. J. Exper. Psych. 88, 67-81.
- Toole, F. E. and B. McA. Sayers. (1965) Inferences of neural activity associated with binaural acoustic images. J. Acoust. Soc. Am. 38, 769-779.
- Treisman, A. M. and G. Geffen. (1968) Selective attention and cerebral dominance in perceiving and responding to speech messages. Quart. J. Exper. Psychol. 20, 139-150.

Left-Ear Advantage for Sounds Characterized by a Rapidly Varying Resonance Frequency

Mark J. Blechner†

ABSTRACT

Nonspeech sounds that could constitute the second-formant transitions of consonant-vowel syllables, with either ascending or descending resonance frequencies, were synthesized. These stimuli were presented monaurally with contralateral noise, and reaction times (RTs) for stimulus identification were measured. RTs were 12.8 msec faster when the stimulus was presented to the left ear than to the right ear, suggesting right-hemisphere involvement in the processing of these stimuli. Ear advantage did not vary systematically with a subject's reported coding strategy for these sounds, whether linguistic or nonlinguistic. These findings suggest that rapid temporal variation is not a sufficient stimulus property to invoke left-hemisphere processing. The roles of other stimulus and coding factors that may invoke lateralized hemispheric processing are considered.

Accounts of the division of labor between the cerebral hemispheres have changed in recent years. In auditory perception, verbal processes have been attributed to the left hemisphere and nonverbal processes to the right hemisphere (Kimura, 1967). This view, however, has not accounted for subsequent data. For example, right-ear advantages (REA) that are presumed to reflect left-hemisphere processes, have been found using nonverbal stimuli, such as sawtooth waves differing in rise time that are musically codable as plucked or bowed violin strings (Blechner, 1976), and rapidly changing tonal sequences (Halperin, Nachshon, and Carmon, 1973). These results, considered along with several studies using visual stimuli (for example, Carmon and Nachshon, 1971; Goldman, Lodge, Hammer, Semmes, and Mishkin, 1968), suggest that rapid temporal variation is a sufficient (although perhaps not necessary) stimulus dimension for invoking the superiority of left-hemisphere processing mechanisms.

According to this view, left hemisphere processes might also be summoned by tasks involving nonlinguistic sounds in which the resonance frequency is rapidly varied. Aside from their rapid temporal variation, such sounds are interesting from another perspective. In a speech context, as part of

---

†Also Yale University.

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

synthetic two-formant consonant-vowel (CV) syllables, they can constitute second-formant transitions that cue the phonemic distinction between voiced stop consonants (Liberman, Delattre, and Cooper, 1952). Yet when isolated from a speech context, the formant transitions resemble the sound of birdsong, and have been called "chirps."

In a variety of experimental paradigms, these chirps yield patterns of results that are distinctively different from other results. For example, in identification and discrimination experiments, stop consonants in CV syllables demonstrate categorical perception, but their isolated second-formant transitions are not perceived categorically (Mattingly, Liberman, Syrdal, and Halwes, 1971). In addition, speeded classification tasks with CV syllables varying in consonant and fundamental frequency have revealed that irrelevant variation in pitch interferes with identification of stop consonants, whereas irrelevant variation in stop consonants does not interfere with pitch identification (Day and Wood, 1972; Wood, 1974, 1975). In contrast to this asymmetric interference, second-formant transitions varying in slope and fundamental frequency produce symmetric interference in the speeded classification task (Wood, 1975, Experiment III).

These results, along with others, have been interpreted as reflective of a dichotomy between auditory and phonetic processes. However, recent experiments with nonspeech stimuli differing in rise time have yielded the same data patterns as speech stimuli in the above-mentioned paradigms, thereby casting doubt on whether the results obtained with speech in these paradigms do in fact reflect a unique phonetic level of processing (Cutting and Rosner, 1974; Blechner, Day, and Cutting, 1976; Blechner, 1976). Nevertheless, comparable results with speech and certain nonspeech sounds leave open the possibility that these paradigms do converge on a levels-of-processing distinction, but that the levels themselves are best characterized by a dimension other than the linguistic-nonlinguistic distinction, such as levels of acoustic complexity or degree of codability. Therefore, it still seems necessary to obtain a complete set of experimental results using a single set of stimuli such as the chirps.

Ear advantage data, which have been considered as another experimental operation converging on the auditory-phonetic distinction (Studdert-Kennedy, Shankweiler, and Pisoni, 1972), have not been reported for chirps. Unpublished studies attempting to use a dichotic identification paradigm with randomly-selected subjects have yielded inconclusive results, largely because of difficulty identifying the chirps<sup>1</sup>.

The present experiment observes whether or not chirps yield a significant ear advantage when presented monaurally with contralateral noise to right-handed subjects, and when reaction time (RT) is measured for stimulus identification. This procedure has yielded a REA for CV syllables

---

<sup>1</sup>Liberman, A. M.: personal communication.



(Springer, 1973). More importantly, Springer's procedure has also yielded a statistically significant REA with musical stimuli differing in rise time (Blechner, 1976), whereas previous dichotic experiments with the same stimuli had yielded null results (see Cutting, Rosner, and Foard, 1975).

## METHOD

### Stimuli

The stimuli were generated on the parallel resonance synthesizer at Haskins Laboratories. They consisted of a frequency-varying pulse-excited resonance. Specified bandwidth was 90 Hz. One stimulus, the rising chirp, varied linearly in frequency from 1232 to 1620 Hz. The other, the falling chirp, varied from 1920 to 1620 Hz. Both chirps had a duration of 50 msec. In a speech context, these two stimuli could cue the distinction between /bae/ and /dae/. However, when presented in isolation, they sound like nonspeech chirps.

These particular resonance frequencies were chosen for another reason. It was decided that the rising and falling chirps should end on the same resonance frequency, since Brady, House, and Stevens (1961) have shown in frequency matching experiments that subjects tend to assign most weight to the final portion of such stimuli, rather than integrating values over the entire stimulus. It was hoped that with the final frequencies equivalent, listeners would be forced to attend to the entire stimulus and its contour in order to distinguish the stimuli accurately.

Noise, presented contralaterally to the stimuli, was also generated by the parallel resonance synthesizer. The noise was 500 msec in duration and its bandpass frequencies were 1232 and 1920 Hz. The noise and both kinds of stimuli were digitized and stored on disc file using the Pulse Code Modulation (PCM) system at Haskins Laboratories. They were reconverted to analog form at the time of tape recording. The absolute levels of the noise and target stimuli, as presented to listeners, were 75 and 68 dB SPL, respectively.

### Tapes

All tapes were prepared using the PCM system. A display tape was prepared to introduce the subjects to the stimuli. The two kinds of stimuli (rising and falling chirps) were played in the same order several times, beginning with three tokens of each item, then two of each, and finally one of each. Two binaural identification tapes were prepared, each with 32 tokens of the chirp stimuli (16 of each) in random order. Four dichotic test tapes were recorded. On one channel of each test tape, 60 tokens of the chirp stimuli were recorded, in random order, with the constraint that every 10 stimuli contained equal numbers of rising and falling chirps. Thus, long runs of any one kind of stimulus were prevented. Sixty tokens of the noise were recorded on the second channel of the tape, with noise onset preceding stimulus onset by 250 msec. The interstimulus interval was 2 sec, measured

between offset and onset of the noise. Four dichotic practice tapes were also prepared. These were identical in design with the test tapes, but contained only 20 stimuli each.

#### Subjects and Apparatus

The 16 subjects included 6 males and 10 females, ranging in age from 18 to 25 years. All were strongly right handed as indicated by the five most reliable criteria found by Annett (1970). All were native speakers of English, and all reported no history of hearing trouble.

The tapes were played on an Ampex AG-500 tape recorder, and the stimuli were presented through matched Telephonics headphones (Model TDH-300Z). Subjects sat in a sound-insulated room and responded with their index finger on either of two telegraph keys mounted on a wooden board. Throughout the experiment, the left key was used for sound #1 responses (rising chirp), and the right key was used for sound #2 responses (falling chirp).

The on-line RT system employed GT-40 and PDP-11/45 computers in tandem. RT measurement was initiated by the onset of the noise, but the 250 msec lead time between the noise and stimulus was automatically subtracted from each RT by the computer program.

#### Procedure

The procedure was nearly identical to that used by Blechner (1976) with different stimuli. Groups of two listeners participated in the experiment. For preliminary training, they listened to the display sequence after being told that the first kind of sound was to be called sound #1 and the second kind sound #2. Subjects were not told until after the experiment that the stimuli typically formed part of speech sounds.

After listening once to the display sequence, subjects were instructed on the mode of response. They then listened to the display sequence twice more, responding first with the left hand and then with the right. This was to insure that they could identify the stimuli correctly. Next, they listened to the binaural identification tapes. Eight of the subjects responded to the first tape with the left hand and the second with the right. For the other eight subjects, the order of responding hands was reversed.

For each individual listener, the chirp stimuli were always presented through the same headphone. Ear of presentation was alternated by having the listener reverse the headset. For eight of the participants, the stimulus was presented through one of the headphones, while for the other eight it was presented through the opposite headphone.

There were four possible hand-ear configurations. The order of these conditions was determined by a balanced Latin square design, yielding four possible orderings that were administered to four subjects each. The four

practice and test tapes, however, were always played in the same order, to prevent any possible confusion between the effects of the random orders and the hand-ear configurations.

Subjects were instructed to respond as quickly and accurately as possible. In the final data analysis, only the last 50 test trials in each block were considered, the first ten functioning as warm-up trials to stabilize performance. The listener, however, was not told that the first ten trials would not count.

After completing all of the listening tasks, the eight last subjects in the experiment<sup>2</sup> were asked to describe in writing, as carefully as possible, "what it was about the sounds that you listened for in order to tell them apart." This introspective task was included to determine (a) what kind of listening strategies were used by the subjects, and (b) whether there was any systematic relation between such strategies and the direction or magnitude of the ear advantage.

## RESULTS AND DISCUSSION

### Left-Ear Advantage

All of the subjects were able to identify the chirp stimuli accurately. In the binaural identification trials, no listener made more than 4.7 percent errors.

For the RT data of the task with contralateral noise, median RT was calculated for each block of test trials for each subject. An analysis of variance was performed on the medians, with order of conditions considered as a between-subject factor, and hand and ear of presentation as within-subject factors. The mean across subjects of individual medians for left and right ear presentation of the stimuli were 686.8 and 699.6 msec, respectively. This 12.8 msec advantage for left-ear presentation was statistically significant,  $F(1,12) = 10.02$ ,  $p < .01$ . No other main effects or interaction terms were significant.

Listeners were very accurate in the stimulus-plus-noise identification task, with the mean error rate only 1.6 percent. An analysis of variance of the accuracy data, identical in design to the analysis of the RT data, showed no significant main effects or interactions.

---

<sup>2</sup>This procedure was not considered until the experiment had been partly run, so that only half of the subjects provided such introspective data.

### Strategy

There was a fair amount of diversity in the strategy subjects reported using in order to distinguish between the two kinds of sound. (See table 1.) Of the eight subjects who provided this kind of data, three reported distinguishing the sounds by means of pitch, suggesting that they attended to the initial or average resonance frequency. Two of the subjects reported using the contour of the formant transition, correctly identifying which was ascending and which was descending. One subject reported idiosyncratic nonlinguistic associations to the rising and falling transitions--a "water drop" and a "buzz-saw," respectively. Surprisingly, two of the subjects reported using a linguistic strategy to distinguish between these supposedly nonspeech sounds: one heard the sounds as different vowels, and the other coded them as nonsense words--"tweet" and "twirt." However, these two subjects did not show any less of a left-ear advantage than did those using nonlinguistic strategies (see Table 1). Of course, these results must be taken with at least three grains of epistemological salt: a) the pitfalls of introspection are well known, and may not reflect the actual psychological processes used by the subject; b) the number of subjects using each strategy is quite small, making judgments of statistical reliability impossible; and c) since the different strategy groups were necessarily determined post hoc, ear-hand configurations could not be balanced between these groups. Thus, one can say that the present data do not support any correlation between linguistic coding strategies and size or magnitude of ear advantage, but they do not rule out such a hypothesis decisively.

### Hemispheric Processes

The present finding of a left-ear advantage for identification of nonspeech sounds that vary rapidly in resonance frequency implicates the right cerebral hemisphere in the processing of these sounds. This result would seem to contradict the hypothesis that processing of rapid temporal variation may be specialized in the left hemisphere. However, in several important respects, the chirps used in this experiment differ from other temporally-varying sounds for which laterality data have been obtained. For example, unlike the CV syllables used by Springer (1973) and the plucked and bowed sounds used by Blechner (1976), the chirps are not readily codable or even familiar to most listeners. It is possible that left-hemisphere processing mechanisms are invoked only by an interaction between temporal variation and codability, regardless of whether the coding of sounds is linguistic or not.

Two other factors should be considered. First, even though the present stimuli were characterized by rapid temporal variation, subjects conceivably could have distinguished between them without processing the rapid variation. They might, for example, have attended to the initial frequency of the chirps. The study of Brady et al (1961) suggests that this is probably not the case, although it is not conclusive.



Second, certain purely acoustic characteristics distinguish the present stimuli from other nonlinguistic sounds that have yielded a REA, and these acoustic characteristics alone may be responsible for the determination of hemispheric specialization. As an analogy, consider the plucked and bowed sounds used by Cutting et al (1975). In a series of identification and discrimination experiments, they found that plucked and bowed sounds yielded the data pattern indicative of categorical perception when the stimulus amplitude decayed gradually over a period of approximately 750 msec. However, when the stimuli were truncated, so that only 250 msec of sound followed the attainment of peak amplitude, the data no longer reflected categorical perception. Unfortunately, no laterality data have yet been reported with such truncated plucked and bowed sounds. Nevertheless, these data of Cutting et al stress the importance of the entire acoustic gestalt as a factor affecting perceptual processes. It is possible that if the chirps were placed in an extended nonspeech context, they might still sound as unfamiliar and uncodable to subjects, but they might nevertheless yield rather different laterality data.

In summary, it appears that rapid temporal variation is not a sufficient stimulus characteristic to yield a right-ear advantage, but currently it is not clear which other factors, such as duration, codability, or listening strategy, may be significant.

#### REFERENCES

- Annett, M. A. (1970) Classification of hand preference by association analysis. Brit. J. Psychol. 61, 303-321.
- Blechner, M. J. (1976) Right-ear advantage for musical stimuli differing in rise time. Haskins Laboratories Status Report on Speech Research SR-47, 63-69.
- Blechner, M. J., R. S. Day, and J. E. Cutting. (1976) Processing two dimensions of nonspeech stimuli: The auditory-phonetic distinction reconsidered. J. Exp. Psychol.: Human Percept. Perform. 2, 257-266.
- Brady, P. T., A. S. House, and K. N. Stevens. (1961) Perception of sounds characterized by a rapidly changing resonant frequency. J. Acoust. Soc. Am. 33, 1357-1362.
- Carmon, A. and I. Nachshon. (1971) Effect of unilateral brain damage on perception of temporal order. Cortex 7, 410-418.
- Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.
- Cutting, J. E., B. S. Rosner, and C. F. F. Foard. (1975) Rise time in nonlinguistic sounds and models of speech perception. Haskins Laboratories Status Report on Speech Research SR-41, 71-94.
- Day, R. S. and C. C. Wood. (1972) Interactions between linguistic and non-linguistic processing. J. Acoust. Soc. Am. 51, 79(A).
- Goldman, P. S., A. Lodge, L. R. Hammer, J. Semmes, and M. Mishkin. (1968) Critical flicker frequency after unilateral temporal lobectomy in man. Neuropsychologia 6, 355-363.
- Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift in ear superiority

- in dichotic listening to temporally patterned nonverbal stimuli. J. Acoust. Soc. Am. 53, 46-50.
- Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.
- Lieberman, A. M., P. C. Delattre, and F. S. Cooper. (1952) The role of selected stimulus variable in the perception of the unvoiced stop consonants. Am. J. Psychol. 65, 497-516.
- Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. G. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.
- Springer, S. R. (1973) Hemispheric specialization for speech opposed by contralateral noise. Percept. Psychophys. 13, 391-393.
- Studdert-Kennedy, M., D. P. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol. 2, 455-466.
- Wood, C. C. (1974) Parallel processing of auditory and phonetic information in speech perception. Percept. Psychophys. 15, 501-508.
- Wood, C. C. (1975) Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. J. Exp. Psychol.: Human Percept. Perform. 104, 1-33.

## An Information-Processing Approach to Speech Perception\*

James A. Cutting† and David B. Pisoni††

### ABSTRACT

The assumptions of information-processing are that: perception occurs over time, is distributed over several stages of analysis, involves several different kinds of memory, and is not infinitely efficient. This paper takes an information processing approach to speech perception. We allow for parallels between perception and production and propose two models of speech perception, one at a macrolevel including all of speech and language, and one at a microlevel including only those processes that seem pertinent to phonetic perception. We review empirical evidence for both models, but concentrate on phonetic perception and relevant experiments with human adults and infants, and with animals. We briefly review some of the continuing controversies within speech perception placed within the information-processing framework. Finally, we suggest some implications of speech research for the school and for the clinic, concentrating on certain aspects of deafness, aphasia, and reading difficulties.

### INTRODUCTION

For most of us, perceiving speech is an effortless and often overlooked task. When engaged in a conversation, for example, we are primarily aware of tracking substance or meaning; the form or sound pattern of what we hear, is "linguistically transparent" (Polanyi, 1964:57), that is, it goes largely unnoticed. The nature of this unnoticed but crucial half of language's dual structure is of particular interest to psychologists, linguists, and engineers, as speech perception is the primary means of picking up information about our culture. The process of converting acoustic information into linguistic message, the underlying structure of that process, and its seemingly unusual design fascinate all those who study speech perception. The nature of this process is also of particular interest to teachers and applied speech scientists. When it goes awry in the young child or adult, it is they who must try to bolster, realign, or circumvent the vocal/auditory

---

\*Presented at the conference "Implications of Basic Speech and Language Research for the School and Clinic", held at Belmont, Maryland, May, 1976. Proceedings ed. by J. F. Kavanagh and W. Strange, to be published by MIT Press, Cambridge, Mass., as Implications of Basic Research in Speech and Language for the School and Clinic (working title).

†Also Wesleyan University, Middletown, Conn.

††Indiana University, Bloomington, Indiana.

system. Our plan in this paper is to take a small step towards mapping some emerging theoretical views of speech perception onto some of the findings in the school and clinic.

We will speak of a process in speech perception, and we mean exactly that. Speech perception is a process because it is not instantaneous; it takes time. We will also speak of a series of stages in this process, organized roughly in a hierarchical fashion. Thus, information of one form will enter a particular stage, be modified or transformed into something new, then enter a new stage, be modified again, and so forth until the linguistic message is understood. Between some of these stages are memory stores, temporary repositories for information that has flowed in. They help break up the dogged linearity of the auditory system. Finally, many of these stages and memory stores have limited capacity. That is, each can hold only so much information of a given kind before it becomes saturated, and information begins to be lost. For those familiar with the zeitgeist of cognitive psychology, these assumptions are easily recognized as hallmarks of the information-processing approach to perception (Broadbent, 1965; Neisser, 1967; Haber, 1969), offspring of information theory and computer modeling within psychology.

We must first establish that these four assumptions are valid--that speech perception is a process, made up of stages, involving memory stores, and that the stages and memories are limited in their capacity or size. Second, the various stages of speech perception will be assembled into a flow diagram, both at a macrolevel including the entire speech/language system, and at a microlevel including only those portions relevant to the lower levels of speech processing. Third, we will present evidence supporting the layout and flow of information to and from each stage. Fourth, we will consider the nature of each phonetic-level stage first in experimental terms, then in terms of ontogenetic development, comparative organization in animals, and possible neurological locus. Fifth, some issues within this information-processing approach will be broached, focussing on particular problems and phenomena at the phonetic level. The final step in this progression will be devoted to possible clinical applications and implications, particularly with regard to what the information-processing scientist would like to know, and what he or she has to offer with regard to the perception of speech sounds.

## SECTION 1: SPEECH PERCEPTION AS INFORMATION PROCESSING

### Speech Perception as a Process

People do not always accept the notion that pattern recognition in general and speech perception in particular, is a process. Malcolm (1971: 386), for example, states that: "When one recognizes a friend on the street there is usually no process of recognition. You see his face in the crowd; you smile at him and say 'Hi, John.' You do not think 'Now where have I seen that face before?'"



Similarly, one might paraphrase and extend Malcolm's statement with regard to speech perception: "You hear his voice in the crowd, speaking to you; you turn and smile and say 'Hi, John.' You do not think: 'Now what did he say and where have I heard that voice before?'"

There are three problems with this kind of refutation of cognitive processing. First, recognition rarely involves conscious subvocal talking to oneself. Second, one need not be aware of a process for processing to occur; just because one is not cognizant of the aspects involved in recognizing the form and content of a friend's voice does not mean that there was no time-course of perception that involved intervening stages. Third, and most important for this discussion, just because recognition is rapid does not imply that it is instantaneous. If we can show that speech perception, for example, takes time, we have strong evidence for a process. In all fairness to Malcolm (1971), there is a kernel of argument underlying his above statement that is the crux of current controversy in speech perception; however, discussion of the controversy centering on invariance will be deferred until Section 5.

Two kinds of evidence support the notion that speech perception is a process. First, there is an upper limit as to how rapidly speech can be understood by the listener. By artificially compressing speech, retaining as many cues as possible, the maximum rate of comprehension is about 400 words per minute (Orr, Friedman, and Williams, 1965; Foulke and Sticht, 1969). At an average of about four phonemes per word, this rate translates conservatively into 30 to 40 msec per phoneme. These high rates are achieved only with considerable practice, only for brief periods of time, and with considerable errors. At faster rates speech melts into a patterned blur (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). One twenty-fifth of a second per phoneme, albeit a very brief period of time, is not infinitesimally brief.

A second line of evidence compatible with the information-processing approach, demonstrates that this time domain, roughly 30-40 msec per phoneme, is correct. If one presents a synthetic consonant-vowel (CV) syllable, such as /ba/ as in bottle, to one ear, and another syllable, /ga/, to the other ear, and further if the onset of /ba/ precedes the onset of /ga/ by about 50 msec, the listener will have considerable difficulty in identifying the first syllable as /ba/ (Studdert-Kennedy, Shankweiler, and Schulman, 1970; Pisoni, 1975a). This difficulty has been attributed to backward masking, a phenomenon in which the identity of a first-arriving item is interfered with by a later-arriving item. The effect of a second stimulus masks, backwards in time, the identity of the first by not allowing pattern recognition processes to operate on it. The domain of this masking effect seems to be about 50-150 msec, an interval that we will consider again in Section 4.

#### Speech Perception is Made up of Stages

To demonstrate that speech perception consists of a series of stages, one could note that the auditory system is made up of component parts: cochlea, cochlear nucleus, trapezoidal body, superior olivary complex,

inferior colliculus, and certainly many cortical elements. We will not use physiological evidence because the existence of separate stages should not rely on physiological and anatomical data. Instead, let us consider some logical requirements of the system.

First, those portions of the central nervous system responsible for speech perception can never be directly "aware" of an auditory signal: they can only be "aware" of, and respond to, neural events transmitted along the auditory pathway. Thus, there must be some stage in the system that transforms auditory signal into neural signal. Second, these portions of the processing system devoted to speech must share the auditory pathway with many other auditorily based systems, much as a single telephone user might share a party line with several other users. Thus, at some level, the neural signal must be general enough to be useful to systems that not only process speech, but other sounds as well, such as music, environmental sounds, infant cries, and the like. Subsequent to this level, it seems likely that the attributes of the neural signal peculiar to speech are transformed a second time into a linguistic description of what has been said, then a third time into meaning (Licklider, 1952).

These three transformations--auditory signal to general neural code, general neural code to phonetic transcription, and phonetic transcription to meaning--may be only a few of those necessary in processing speech. Indeed, as we will see later there are probably many more. Liberman, Mattingly, and Turvey (1972) have estimated that the transformation of a reasonably intelligible, acoustic signal into a phonetic representation of speech--involving two of the transformations just mentioned--is equivalent to transforming a 40,000 bit-per-second signal into a 40 bit-per-second signal. This rapid thousandfold reduction of information results in a coded form of speech that is suitable for even further reduction by coding into meaningful linguistic units. Such magnitude of information-reduction is requisite in speech perception. Moreover, such feats could only be accomplished by having the signal pass through, and be transformed by, several different processing stages.

The exact number of stages needed for the processing of speech is not known, although there is common agreement about their general layout in a flowchart. We are all familiar with the convention of drawing boxes and lines to represent stages and the flow of information between them. That enterprise will be repeated here. Drawing boxes and connecting them with lines is, as Roger Brown (1973:4) might say, "an odd interest, dependent [we] suspect, on some rather kinky gene which, fortunately for our species, is not very widely distributed in the population." Nevertheless, fruitful insights and testable hypotheses arise from such a venture. In formulating these pencil-and-paper systems, however, one must be careful. The information-processing scientist must be selective but not overly economical in the number of stages postulated in the system. On the one hand, he or she must try to avoid such atrocities as Whorf's (1940) tongue-in-cheek fifteen-stage process for translating English into French; yet on the other, he cannot postulate too few stages. Otherwise he regresses ultimately to Malcolm's (1971) one-stage direct processing view of perception.

### Memory Stores for Speech Perception

Most of us have experienced the following embarrassing situation. Imagine yourself sitting across the table from a friend. Both of you are absorbed in different activities, but intermittently you talk to one another. A period of silence goes by and suddenly, your friend asks you a question. About the time she finishes you say: "Sorry, what did you say?" Distressingly, almost before you finish, and certainly before she has a chance to repeat the question, you already know what it was. From where did the question re-emerge? The answer must be that it had been stored unused in some kind of memory until the general speech processing system and consciousness accessed it.

This kind of anecdote demonstrates that speech perception can use memory. We, and Robert Crowder (in press) hope to demonstrate that speech perception requires the use of memory. Let us preface this discussion with a brief note on why more than one kind of memory is necessary in an information-processing approach to speech perception. Quite aside from the anticipated inclusion of echoic memory, short-term (or primary) memory, lexical memory, and semantic memory within the general speech/language system, one must remember that speech perception is very fast, yet is made up of several stages of processing. Add to this the fact that speech is a dynamic signal in which interrelations among spectral parts are constantly changing, and changing at a constantly varying rate. Some portions of the signal may be very easy to process, perhaps the relatively steady-state portions of stressed vowels, while others, perhaps the bursts and rapid transitions of stop consonants, are more difficult. These variations must be accounted for and handled. The only plausible manner in which this can be done in "real time"--that is, without stepping outside the natural context and artificially slowing down the signal so that the more recalcitrant aspects of the signal become amenable to analysis--is through the copious use of a series of memory stores or buffers, whose contents are constantly updated and overwritten by subsequent information. Moreover, the contents of these stores must be accessible to a number of stages in the system. We will try to develop the interrelations of these buffers and stages in Section 2.

### Limited Capacity of Nervous System

Although the central nervous system may be made up of billions of cells and quite possibly many millions of them are devoted exclusively to speech and language, the resources of this system are far from unlimited. Consider first the memories. Echoic memory, often thought of as a kind of an audio tapeloop that is constantly rerecorded, lasts only about one or two seconds. Thus, we find a temporal limit to the amount of information that can be stored in a fairly raw form. If we wish to "play back" a speech sample that we have just heard, we must do it quickly, otherwise its relatively unencoded form will be gone forever (Crowder, 1971; Pisoni, 1973; Darwin and Baddeley, 1974).

Short-term memory is the conscious, verbal memory so well studied by psychologists. It is the memory that fails when one forgets a telephone

number between directory look-up and dialing. It seems not to be limited so much by time as by amount and content. Seven (Miller, 1956), or more likely five (Broadbent, 1975) unrelated items are about all this memory can hold and recycle in relatively crude form for further analysis. The items themselves can be syllables, words, or even multiword units, but it seems that they cannot be sentences of any length, and certainly not paragraphs. Thus, if a linguistic message is to be comprehended, its gist must be quickly abstracted and recoded or run the risk of never being fully processed [see Bransford and Nitsch (in press)].

It would seem that echoic memory is a neurologically expensive memory, whereas short-term memory is less so. If, in fidelity, echoic memory approaches the quality of tape recording, it will require an approximation of the 40,000 bit-per-second storage mentioned earlier (see also Norman, 1972). While the early stages of auditory analysis could certainly handle this load, one can easily see why humans were not engineered to have an indefinitely long echoic memory; perhaps billions of brain cells would need to be involved. Short-term memory, on the other hand, is likely to be of the 40 bit-per-second variety, and better adapted to hold the more highly coded linguistic message. It is limited in its capacity nonetheless, whether by neurological design or by evolutionary caprice. Only lexical memory (the dictionary-in-the-head) and semantic memory (that used for comprehension) are thought to be functionally unlimited in their capacity. Their lack of limits would make them expensive neurologically, and in a pneumatic fashion, force capacity constraints onto other parts of the system.

The most important capacity limitations, however, are not those placed on memories; but those placed on the entire system by attentional processes. It appears that we can pay attention to only one thing at a given time (Broadbent, 1958, 1971). Thus, from all the inputs from all perceptual systems, only one source of information can ride high within consciousness, or in an information-processing analysis of attention, only one source of input can survive attentional selection. The locus of this squeeze on various inputs has been the source of controversy for twenty years, but it now appears that attentional selection occurs quite late in the system after perceptual processing (see for example, Shiffrin, Pisoni and Cataneda-Mendez, 1974). This fact is important in speech perception because, as noted in the introduction, we often are aware of attending to meaning of discourse without direct awareness of sound pattern. Attentional constraints would appear to play a role in the "transparency" of sound. Moreover, if attention is subsequent to perception, the representation of speech in the what-did-you-say example cited previously is likely to be in a more highly coded form than that of echoic memory; the message was processed and awaited only attentional focus. Full awareness of the sound pattern of speech may be possible only when we have disengaged ourselves from meaning, such as when we listen to a conversation spoken in a language that we do not understand or to the first babblings of a young child.



## SECTION 2: AN INFORMATION-PROCESSING MODEL OF SPEECH PERCEPTION

### Production and Perception Together

Before putting these assumptions together into an information-processing model of speech perception, it is necessary first to establish a conceptual framework for it. De Cordonoy first postulated a connection between the perception and production of speech in the seventeenth century, but it is only since Lashley that this notion has been taken seriously by psychologists. Lashley (1951:120) appealed to parsimony: "The processes of comprehension and production of speech have too much in common to depend on wholly different mechanisms." Until recently, however, we have had little more than parsimony and implausibility of alternatives to substantiate this statement (more on this at the end of Sections 4 and 5).

Some of the processes thought to be held in common between perception and production are shown schematically in Figure 1. In producing speech, for example, we start with some conceptual representation, coded in "mentalese" (Fodor, Bever, and Garrett, 1974), and move through a series of at least four other stages until we reach the acoustic structure of speech. Thus, at the two ends of the process we have meaning and sound. Between them are stages of deep structure, surface structure, and phonetic structure, and a series of transformation processes--semantics, syntax, phonology, and speech. It should be noted that this system can easily be elaborated, either at the conceptual end (see Fodor, Bever, and Garrett, 1974:391) or at the speech end (see Cooper, 1972:34). The importance of this display, however, is that we can reverse the arrows for the speech production process to achieve a fairly accurate conceptualization of speech perception.

### Parallel Processes

What is not achieved, however, in this conceptual display of speech production and perception is the impression of multiple interrelations between, and simultaneity of operations within, these stages. While the stages are likely to be serial in some sense--after all speech is inherently a temporal process--they must also be parallel. This is not quite like having one's cake and eating it too. Decisions at one level can be made on the basis of preliminary information sent up or down the system from another level like a progress report, rather than each stage waiting for ultimate decisions, as in a final report. Again, we need not be aware that these intermediate decisions are being made; we need only be aware of the ultimate outcome. In such a dynamic system there must be careful executive monitoring of parallel processes so that each stage does not act on misinformation. Improper monitoring for misinformation may be responsible for occasional metatheses, or spoonerisms, in speech production (MacKay, 1970; Fromkin, 1971) and a tendency to miss errors of pronunciation in speech perception (Cole, 1973; Marslen-Wilson, 1975). Thus, while the hierarchical representation of the stages of processing represented in Figure 1 is probably not wrong, it certainly can be misleading. Perhaps a better organization, at least from an information-processing point of view, is a more hetarchical one proposed by us elsewhere (Pisoni, 1975b, in press b),

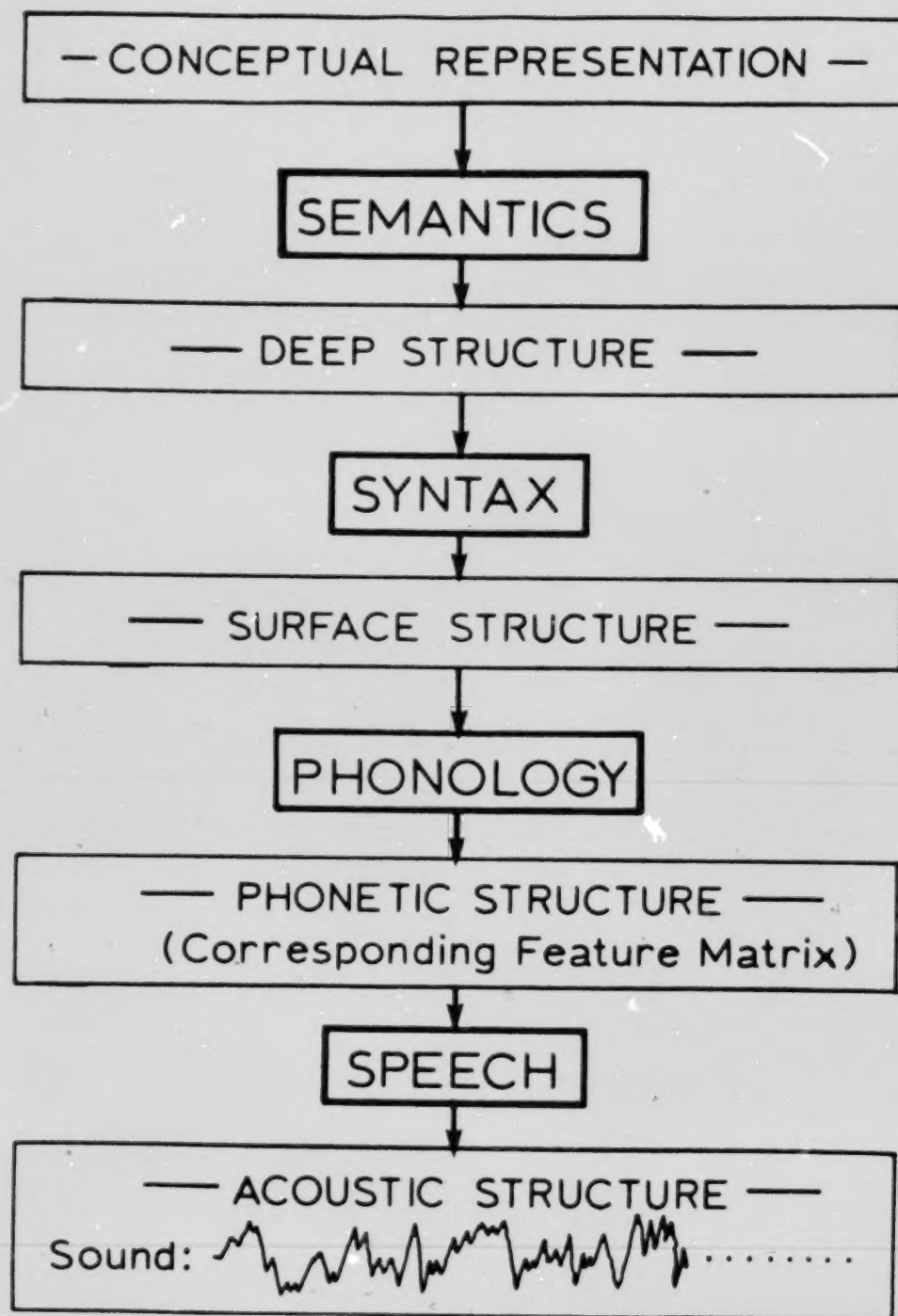


Figure 1: Serial organization of some stages of speech production. Arrows can be reversed for an approximation of the process in speech perception (adapted from Liberman, 1970).

and shown in Figure 2.

This is a macrolevel model of speech perception, including in it the entire speech/language system. It allows in real-time operation, the simultaneous functioning of phonetic, phonological, lexical, syntactic, and semantic processes to derive a linguistic representation of a sentence to-be-perceived. Its advantage is that it is fundamentally a dynamic approach to speech and language perception rather than a template-matching, taxonomic one (Massaro, 1975); it allows language to act as a "support system" for speech perception (Bransford and Nitsch, in press).

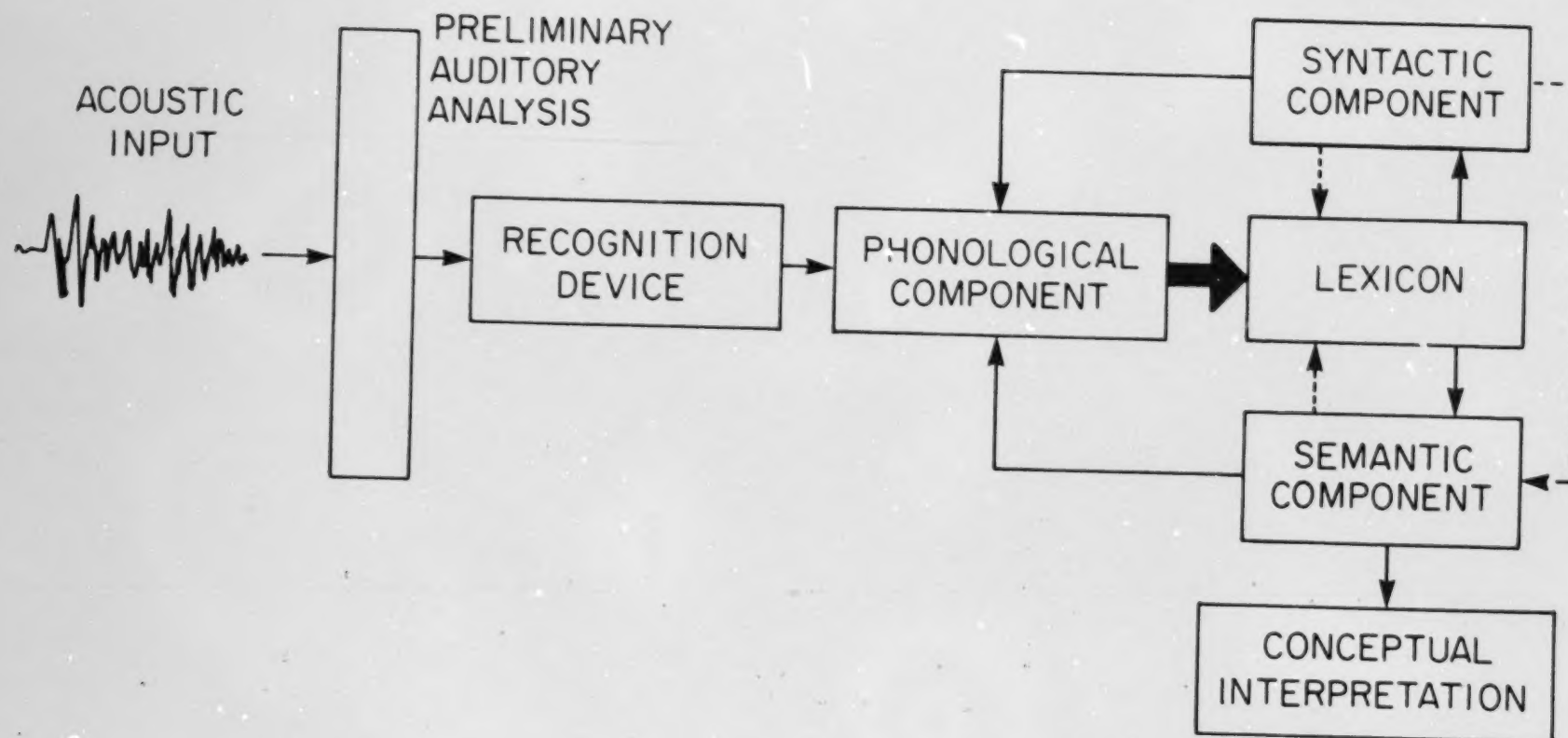
Since the major portion of this paper is directed towards speech perception rather than language perception, we will elaborate the recognition device, shown in Figure 2, and describe a microlevel model. This device handles only the phonetic recognition process, and it is shown in Figure 3 as Pisoni and Sawusch (1975) first conceived of it.

The auditory input enters the block-diagrammed system and undergoes a series of transformations. First it enters a stage called preliminary auditory analysis, where acoustic energy is transformed into neural energy, preserving to a large degree the time, frequency, and intensity relations among the components of the signal. The information then enters a sensory information store. Many information-processing models call this stage preperceptual auditory storage (Neisser, 1967; Massaro, 1975; Massaro and Cohen, 1975). This storage is thought to be very brief--on the order of 50 to 250 msec--and is thought to account for, among many other phenomena, the backward masking results discussed in the introduction. The auditory system does not appear to be able to resolve separate information within a smaller time domain than 20 to 40 msec (Hirsh, 1974; Stevens and Klatt, 1974). This smearing, or integration of the temporal signal occurs in preliminary auditory analysis. Preperceptual auditory storage appears to reflect a "time window," or perceptual moment (Allport, 1968) that travels along the acoustic signal. The window has--not sharp--but graded edges that may extend the integrating field to as much as 100 msec or more in optimal circumstances.

Information is further transmitted to the recognition device and undergoes a series of at least four stages of analysis. These stages, and the previous registration of the acoustic signal within the sensory information store, are mandatory, and not under the conscious control of the listener; they occur more or less automatically.

Within the recognition device are stages of auditory feature analysis and phonetic feature analysis, a phonetic feature buffer, and a mixer in which the coded signal undergoes phonetic feature recombination. In stage 1 of the recognition device, auditory properties of the speech signal are recognized, in parallel, by a whole system of units whose sole job is to parse the incoming information looking for prominent auditory attributes. Following Stevens (1975), some attributes that might be included are: (a) the presence or absence of rapid change in the spectrum--information that can aid in the recognition of consonants versus vowels; (b) the direction, extent, and duration of change within a portion of the spectrum--information

Figure 2: Functional organization of the components of the speech perception system.





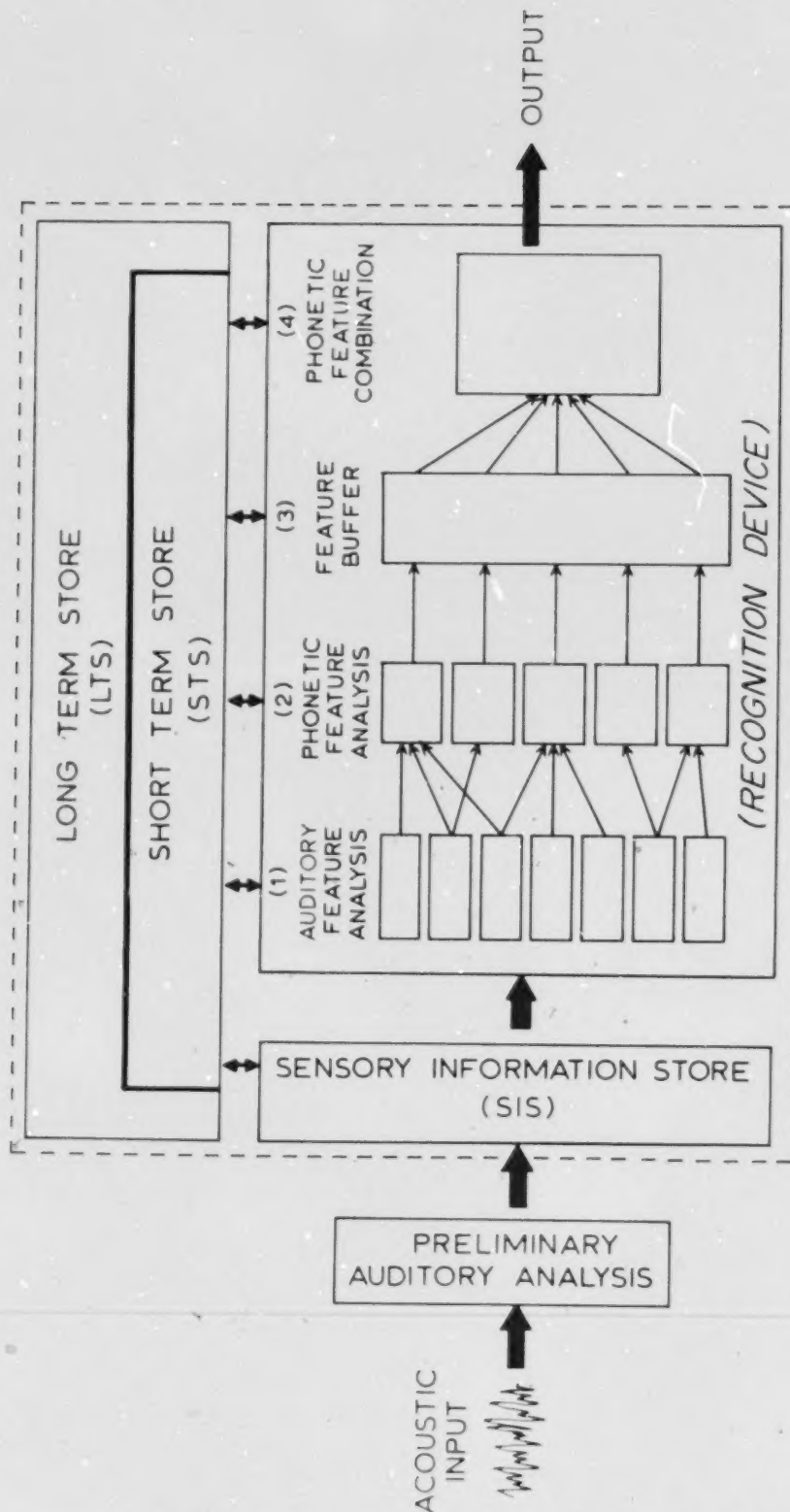


Figure 3: Functional organization of the phonetic recognition component of the speech perception system (from Pisoni and Sawusch, 1975).

that can aid in distinguishing consonants from one another; (c) the frequency range, duration, and intensity of noise--information that can aid in the distinguishing of fricatives from one another (Hughes and Halle, 1956; Gerstman, 1957); and (d) the relative onset of periodic and aperiodic portions of the signal (Lisker and Abramson, 1964; Lisker, 1975), to name just a few acoustic cues. The output of auditory feature analysis is some combination of the possible cues in the signal, which is then sent on to the next stage of processing (see Sawusch, 1976).

Stage 2 concerns phonetic feature analysis, where a complex of decision rules maps the multiple auditory properties onto phonetic features. It is within this stage that neural signal becomes language, and it is about this stage that we will have most to say in later sections. This stage can be presumed to have access to, or knowledge of, articulation constraints of the vocal tract. The output of this many-to-one and one-to-many mapping is a set of abstract phonetic features, sent on to the third stage.

Stage 3, the feature buffer, is a form of memory not previously discussed. It is simply a holding bin that preserves the phonetic feature composition of the particular syllable being processed. A feature buffer is needed here because we cannot assume that all phonetic features are processed at the same rate. Moreover, some memory is needed to preserve and maintain phonetic feature information independently for subsequent linguistic processing, particularly for phonological processing, while prior stages begin to process later-arriving material. This holding bin is intimately related to short-term memory.

Stage 4, the final stage within the recognition device, is a mixer for the recombination of phonetic features. That is, it is at this stage where clusters of phonetic features are assembled into a phonetic string. This time-lagged distinctive feature matrix is sent on to higher levels of linguistic processing that modify and extend it through phonological, lexical, syntactic, and semantic analyses.

Before being sent upstream, however, information from any of these four stages and from the sensory information store can be placed into short-term memory, where the listener first has control over it and can selectively rehearse, encode, or make decisions about it. Long-term memory is also assumed to be accessed during the recognition process, and consistent with recent accounts of their relationship, short-term memory is thought to be simply that portion of long-term memory temporarily activated (Bjork, 1975; Shiffrin, 1975). We consider long-term memory as including episodic memory (Tulving, 1972), which stores "episodes" of personal history according to spatial and temporal tags, and more importantly, semantic and lexical memory (Collins and Quillian, 1969; Miller, 1972, in press), which are thought to be those portions of long-term memory "necessary for the use of language. [They are the] mental thesaurus, organized knowledge a person possesses about words and other verbal symbols, their meanings and referents, about relations among them, about rules, formulas, and algorithms for the manipulation of these symbols, concepts, and relations" (Tulving, 1972:386). While not overtly marked in Figure 3, short- and long-term memory must also be able to access

the higher levels of the speech/language system diagrammed in Figure 2; after all, the lexicon, syntactic rules, and meaning are part of long-term memory (LTM).

We regard the models in Figures 2 and 3 as preliminary, and suppose that a number of revisions will be needed. Nevertheless, these current forms provide a convenient framework in which to place many of the results of research in speech perception that have occurred over the past twenty-five years. Moreover, these models should suggest the kinds of things we need to know about speech perception in the normal population, and perhaps those things that experiments with listeners belonging to special populations might be able to tell us. First, let us consider evidence that supports the general information-processing model.

### SECTION 3: EVIDENCE SUPPORTING THE INFORMATION-PROCESSING MODEL

#### Language Contributions to Speech Perception

First, consider the macrolevel model presented in Figure 2. Evidence supporting the necessary inclusion of higher-order components in the process of speech perception comes from many sources. Perhaps the most direct proof comes from the many studies of the perception of speech under noise. Here, "redundancy" of the speech pattern makes otherwise unintelligible sections of speech considerably more intelligible. Within the context of this information-processing approach, we can attribute this redundancy gain to contributions of higher levels of processing in the speech-language system. Bransford and Nitsch (in press), for example, discuss the contribution of a conceptual component to the perception of speech under noise.

Two components of the speech/language process are syntax, which is good word order, and semantics, which relates to overall meaning. Both these components can be seen working together in the results of Miller, Heise, and Lichten (1951) and Miller (1962). The intelligibility of isolated words at a signal-to-noise ratio of 0 decibels (dB)--when the signal and noise have the same amplitude--was only about 40 percent. Those same words, however, when they appeared in sentences, were intelligible nearly 70 percent of the time. This 30 percent increment in performance can be attributed to syntactic and semantic components shown in Figure 2. Thus, when the speech signal is impoverished, the higher-level components of the system can help tease out the important features from the background of noise. How exactly this is done is not known, but many have suspected, and we agree, that linguistic context reduces the number of alternatives. A similar effect can be seen in the results of Pollack and Pickett (1963, 1964), where the intelligibility of excerpts from conversation are shown to be directly related to the duration of the excerpts and to the number of words contained in them.

Syntactic and semantic components can also work separately. Again burying speech in noise, Miller and Isard (1963) found that ungrammatical, largely meaningless strings of words were more difficult to repeat than grammatical strings that were equally meaningless. These sentences, in turn, were more difficult to repeat than grammatical strings whose meaning was easy

to obtain. Thus, syntax improves speech perception, and the addition of an easy-to-interpret semantic component improves it even further.

Lexical memory, the memory concerning specific meanings of words, also plays an important role in perceiving speech. Varying signal-to-noise ratios, Howes (1957) and Savin (1963) have shown that word recognition for common words is very different from that for rare words; common words are recognized and identified when buried in noise 20 dB more intense than the corresponding threshold level for rare words. This phenomenon is presumably attributable to the fact that common words are highly used and most easily accessible in lexical memory. Rare words, on the other hand, subsist in musty seldom-visited corners of the lexicon. In terms of information theory, a common word is a much more plausible alternative than a rare word in an ambiguous situation. Plausability can be thought of as a psychologically weighted value related to frequency of use, attached to each word in the lexicon. A related concept is that of predictability and the number of alternatives in an ambiguous situation. Miller, Heise and Lichten (1951) varied the set size of words to be played under noisy conditions and found that, for example, at a signal-to-noise ratio of -12 dB, words from a set size of 256 items were about 10 percent recognizable, those from a set size of 32 items were about 40 percent recognizable, and those from a set size of four were about 70 percent recognizable. Clearly, the number of possible alternatives in a situation of signal in noise, and the psychological weights attached to those alternatives, play a vital role in speech perception.

The phonological stage is where the phoneme first appears; phonological cues can also be a great aid. This component provides information about the sound structure of a given language, and this information is imposed on the phonetic feature-matrix in order to derive a phonological matrix. Aspects of this component are both universal and language specific. General syllabic form and information about intonation contour or prosody, are also processed at this stage.

Stress patterns, which is one aspect of prosody and a speech variable processed by the phonological component, can be a potent cue in word identification. Kozlowski<sup>1</sup> has used this fact as a tool in the study of tip-of-the-tongue (TOT) phenomenon (Brown and McNeill, 1966). The TOT state is a tantalizing mental condition in which the individual searches for a word, knows it is there, but cannot obtain it from lexical memory. Certain attributes of the word, however, can be accessed, such as syllabic structure and initial and final letters. Kozlowski employed this information by presenting his subjects with definitions of rare words thought likely to be in a college student's recognition vocabulary, but not in their active vocabulary. When his subjects declared themselves to be in the TOT state he presented them with an auditory cue. This cue was a severely low-pass-filtered version of the target word, where the filtering essentially removed

<sup>1</sup>L. T. Kozlowski. Poets, nonpoets, the tip of the tongue phenomenon, and the perception of poetry. (Manuscript submitted for publication).



from the speech signal everything except fundamental frequency, some gross aspects of intensity envelope, and general syllabic structure--hence only the stress pattern of the target word remained. The end result sounded rather like an individual talking through a pillow. Kozlowski found this cue potent enough so that his listeners could immediately perceive the target word in 35 percent of all TOT states. False cues, filtered renditions of words that were not the correct answer, provided only 15 percent "recognition" of the TOT word, a rate that may reflect spontaneous remission of the TOT state. Clearly, phonological cues are important to the perception of speech; knowing the substance of a word but not its form, one can often perceive a word when only its stress and syllabic patterns are available.

Before turning to the phonetic level model of Figure 3, we might add one more extraphonetic component that aids the perception of speech. This component, not shown in Figure 2, is often overlooked by the nonapplied speech researcher--the facial display of the speaker (O'Neill, 1954; Sumbly and Pollack, 1954; Erber, 1969, 1975). Erber (1969), for example, notes that listeners with normal hearing who try to understand words spoken under intense noise (signal-to-noise ratios of -20 dB) may not be able to identify any words. Give to these same listeners under the same noise conditions, the same words but now with the opportunity to observe the speaker, and their identification improves to as much to 60 percent. Not surprisingly, much of this gain is attributable to visual access to cues of place of articulation (Binnie, Montgomery and Jackson, 1974), a feature that is not very resistant to noise-embedding (Miller and Nicely, 1955), but one that carries a wealth of linguistic impact. As Erber (1975) notes, for the hearing-impaired individual, most of speech perception is a task very similar to this latter experimental situation.

Note that the particular layout of stages in Figure 2 allows nearly maximum interaction between phonological, lexical, syntactic, and semantic components. We choose this plan, in part, to hedge our bets. The role of the lexicon has created a sticky issue among both linguists (Fodor, Bever, and Garrett, 1974) and psychologists (Miller, in press). One aspect of the issue is the question of where in the general scheme of syntax and semantics, do individual isolated words that occur in the process of generating and perceiving sentences, fit? There appears to be no clear answer as yet, nor do we have one, and so we have placed the lexicon near the middle of the macrolevel model. Current views of transformational grammar make similar placements of the lexical component (Bresnan, 1976). After all, the lexicon and other higher-level stages are part of long-term memory.

One may wonder: why have we bothered to go through extensive explanations concerning aspects of language that overtly have nothing to do with speech perception--at least as that subdiscipline has come to be known? In addition, one may ask: Why have we bothered with all this evidence under noise? The answer to the first question is twofold. First, an informationprocessing account of a system as complex as speech perception demands thoroughness. We must consider the whole system; without a holistic approach a proper phonemic description is not possible (Chomsky, 1964). Second, language perception is a dynamic, whirling process for which speech

perception in most of us, is the linchpin. Speech and language are not easily divisible in a working system, just as a wheel and its hub are not easily divided in a moving vehicle. To consider speech without regard for the higher processes of language is, if not an empty pursuit, certainly one which will mislead both the basic and the applied researcher. The answer to the second question is similar in tenor. Speech perception under conditions involving some level of background noise--be it patterned, white, or shaped--is speech perception as we do it every day. Speech and noise are as natural in combination as speech and language.

#### Phonetic Contributions to Speech Perception

For speech perception--or phonetic perception--we will consider results that are fairly recent. Many experimental outcomes appear to converge on a model such as that shown in Figure 3 (see for example, Wood, Goff, and Day, 1971; Day and Wood, 1972; Studdert-Kennedy, Shankweiler, and Pisoni, 1972; Cutting, 1974, 1976; Wood, 1974, 1975). In general, these results point to the facts that: (a) speech is treated by the listener as a multidimensional display, some attributes of which are auditory with little bearing on language, and some which are phonetic and integral to the speech code; (b) these auditory and phonetic attributes appear to be coded differently in memory and may be established in different parts of the central nervous system; and (c) auditory processes are logically prior to phonetic processes, but in most situations the two types of processes go on in parallel.

#### SECTION 4: NATURE OF THE STAGES IN PHONETIC PROCESSING

##### Experimental Evidence With Normal Adults

After preliminary auditory analysis, the signal is transmitted to a sensory information store. Much of this information is then sent to echoic memory (Neisser, 1967; Crowder and Morton, 1969) which we view as a component of short-term memory. Placing this storage in the system early has the advantage that the listener can "postpone classification of some items momentarily, recheck his categorization of others, and, generally, transcend the strict pacing" imposed on him by the temporally linear aspect of audition (Crowder, 1972:254-255). What resides within this memory store is an auditory code of the input, particularly well suited to prosodic features and certain speech segments. It appears, however, that all speech sounds are not equally suited for such a code. Stop consonants, for example, appear to be considerably less amenable to such storage than vowels (Crowder, 1971; Pisoni, 1973). Moreover, the differences between the two phoneme classes does not appear to be related to phonetic coding, but rather to the fact that rapidly moving transients such as those found in stop consonants cannot be laid down in auditorily coded form as easily as the more steady-state attributes of vowels (Darwin and Baddeley, 1974). While certain auditory properties of stop consonants can be accessed by short-term memory and consciousness (Barclay, 1972), the more typical situation is one in which only the form of the vowels (and perhaps the fricatives) can be extracted from sensory information store. Thus, much of the raw signal may bypass the echoic portion of short-term memory and be transmitted directly from sensory

information store to auditory feature analysis. This "bypassing" is a signal-dependent process: what can be stored in echoic memory is stored, what cannot be stored is not, and the rules for what is and is not stored are dependent on the nature of the particular properties of the signal.

Next in processing, is auditory feature analysis--a stage which has recently assumed a larger role in accounts of speech perception (Stevens, 1972, 1975). For vowels and other relatively easy-to-process segments, this stage may not be vital. An auditory husk may be available in echoic memory for up to two seconds before the signal need be transformed into a more parsimonious phonetic-feature code. In fact, steady-state vowels might be directly coded into phonetic form from the echoic portion of short-term memory (STM). For consonants, on the other hand, the auditory-feature stage would appear to be essential.

Perhaps the best experimental evidence supporting the existence of this property-detection stage comes from a paradigm recently imported from vision research (Blakemore and Campbell, 1969). It is called selective adaptation. Although its original purpose was to rally support for a direct, phonetic-feature-detection model of speech perception (Eimas and Corbit, 1973; Eimas, Cooper, and Corbit, 1973), recent evidence using this paradigm supports an auditory property view (see Cooper, 1975, for a review; J. L. Miller, 1975; Pisoni and Tash, 1975; Tartter and Eimas, 1975; J. L. Miller and Eimas, 1976; Cutting, Rosner, and Foard, (in press)). In the adaptation situation, the listener is presented with dozens, perhaps even hundreds, of tokens of the same utterance and then asked to identify members of an array of stimuli. Results show that previously ambiguous items in the array--those at the boundary between two categories of stimuli--are afterwards identified as unambiguous exemplars of the stimulus category opposite to that which has been adapted. This result is important because it suggests how speech might be perceived from opponent-process binary devices relatively early in the information-processing system. It is also complex, and thus it behooves us to first discuss two other phenomena--categorical perception and chromatic afterimages.

Categorical perception is a peculiar, nonlinear mode of perceiving typically associated with stop consonants. Given an array of seven stimuli from /ba/-to-/da/, shown schematically in the top left-hand side of Figure 4, randomized and presented many times, the listener usually identifies stimuli 1 through 3 as /ba/, and stimuli 5 through 7 as /da/. Stimulus 4 is identified as /ba/ about 40 percent of the time, and /da/ about 60 percent of the time. When these items are discriminated, the listener finds it very difficult to tell the difference between stimuli 1 and 3, for example, or between stimuli 5 and 7, but she has no difficulty in discriminating stimulus 3 from stimulus 5. This set of results is interesting because the seven stimuli in this /ba/-to-/da/ array differ from one another in equal acoustic increments. That is, in terms of the amount of difference in start frequency of their second formant-transitions, stimuli 3 and 5 are no more different than stimuli 1 and 3 or stimuli 5 and 7. Typical results of the identification and discrimination tasks are shown in the lower right pane of Figure 4 (see Liberman, Harris, Hoffman, and Griffith, 1957; Studdert-

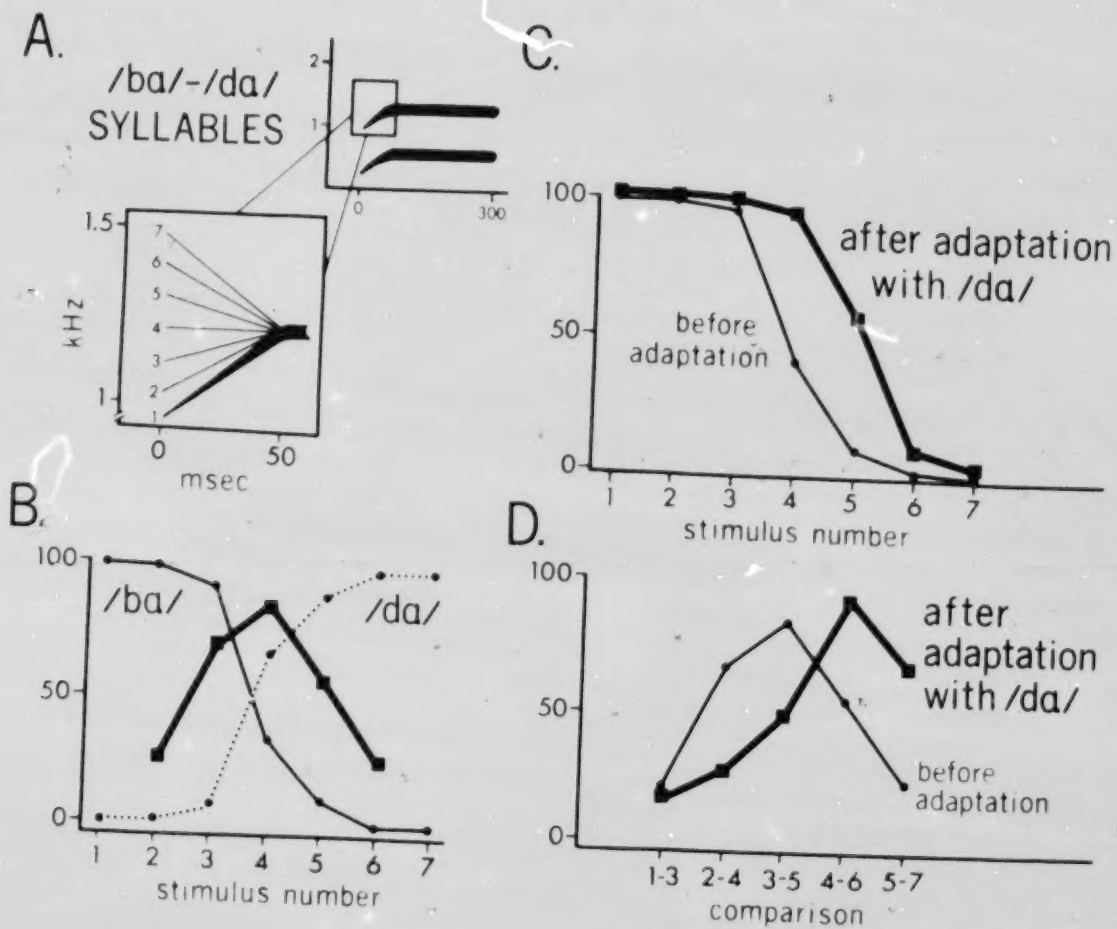


Figure 4: (a) Schematic spectrograms of an array of stimuli from /ba/ to /da/, (b) identification and discrimination functions for that array, and (c & d) those functions before and after adaptation with stimulus 7, /da/.



Kennedy, Liberman, Harris, and Cooper, 1970; Pisoni, 1973).

Now let us make a temporary jump to the perception of color. A well-known phenomenon here is the chromatic afterimage (not the one studied by Blakemore and Campbell, 1969). If a viewer stares at a patch of blue for fifteen to thirty seconds, then stares at a blank white wall illuminated to the same degree, he sees a patch of yellow whose contour conforms to the original blue patch. Blue is, at some level of analysis, the opposite color from yellow; white is in between. Staring at the blue patch fatigues the blue receptors in the retina, and sets up a temporary imbalance in the opponent-processing system for color. Presented with a neutral stimulus--a white wall--the viewer then sees a color that is unambiguously yellow--the opposite from blue.

In the speech domain, /ba/ can be thought to be the opposite from /da/. Since the listener cannot "stare at" or even listen to, a speech syllable for any length of time--two seconds may be the outside duration of such an echo, and stop consonants simply do not echo--a syllable such as /da/ must be presented over and over again to replenish the auditory trace. When this is done, both the identification and the discrimination functions for the set of stimuli are temporarily shifted towards /da/, as shown in the right-hand panels of Figure 4. The analogy between the effect and chromatic afterimage breaks down at this point. The adaptation effect for speech syllables is not in the cochlea (the counterpart to the retina), but farther along in the processing system, and at least partly beyond that point in the auditory pathway where the two ears converge (Eimas, Cooper, and Corbit, 1973).

Note, however, that the /ba/-to-/da/ stimuli differ in what Whitfield (1965) has called "auditory edge": those stimuli identified as /ba/ have a rising second-formant transition, whereas those identified as /da/ have a falling second-formant transition. It is a bit unusual that the boundary between two stop consonants in consonant-vowel (CV) syllables falls at a point where there is no second-formant transition, but this case is not unique and probably many other such boundaries can be seen as a variation on this theme. It would appear that in the adaptation situation where /da/ is the adapting stimulus, stimulus 5 rather than stimulus 4, is now "perceived" to have a zero-sloped transition at stage 1 of our model. By this account, it is the auditory feature-analysis stage of our model that has been temporarily affected. Arguments for this interpretation instead of a direct, phonetic feature-detector interpretation are complex and the reader is referred to Cooper and Blumstein (1974), Pisoni and Tash (1975), Tartter and Eimas (1975), J. L. Miller (1975), Cutting (in press), and Cutting, Rosner, and Foard (in press). In summary, however, the assumption that adaptation effects were due to "fatigue" of phonetic feature-detectors appears to have arisen from the unnoticed pun between the phonetic features (redistinctive features of Jakobsen, Fant, and Halle, 1951) and the term feature (property) detectors, as the concept was borrowed from the vision literature.

Auditorily coded properties from stage 2 are mapped in a many-to-one and one-to-many fashion onto phonetic features in stage 3. A many-to-one relationship is necessary because a phonetic feature value, such as the

voicelessness of /p/ in English, can be cued in many different ways. In syllable-initial position, for example, this feature value can be cued by cutback in the first-formant transition (Liberman, Delattre, and Cooper, 1958), and by delay in voice-onset time (VOT) (Lisker and Abramson, 1967). In intervocalic position, this distinction can be cued by the silent interval between the offset of the previous syllable and the onset of the target syllable (Liberman, Harris, Eimas, Lisker, and Bastian, 1961). In syllable-final position it can be cued by the duration of the previous vowel (Raphael, 1972). A one-to-many relationship is also needed since, for example, a single burst can cue different stop consonants (Liberman, Delattre, and Cooper, 1952; Schatz, 1954). The phonological component further maps phonetic features onto phonemes in a second type of many-to-one and one-to-many fashion, not shown in Figures 2 and 3.

Support for the existence of a feature buffer (stage 3 of the recognition device) and a store in which these features are recombined (stage 4) comes from dichotic listening tasks. If /ba/ is presented to one ear and /ta/ to the other ear simultaneously, the listener often reports hearing a syllable not presented, and most often that syllable is /da/ or /pa/ (Halwes, 1969; Cutting, 1976). Consider the response /da/. It appears that the voicing feature value for /ba/ is perceptually combined with the place-of-articulation feature value for /ta/, with the result being /da/. This is an example of perceptual synthesis of a new syllable from the phonetic feature values of the stop consonants presented to opposite ears. This combination appears to be phonetic because variation in the carrier vowel of the syllables appears to have little effect on the frequency of such "blends" (Studdert-Kennedy, Shankweiler, and Pisoni, 1972); /bi-/tu/ pairs appear to yield as many /d/ (and /p/) fusion responses as /bi-/ti/ pairs. These fusions, or phonetic feature value combinations, appear to occur in stage 4 of Figure 3. Stage 3, the feature buffer, on a /bi-/ti/ trial, would probably contain all phonetic feature values of the two stop consonants--voiced and voiceless manners of production and labial and alveolar places of articulation. Since a stop consonant cannot simultaneously be voiced and voiceless, or labial and alveolar, only one feature value of each can be combined in a response.

#### Experimental Evidence With Young Infants

The one- to four-month-old infant does not possess any lexical, syntactic, or semantic processes, nor phonological processes that match in any degree those of adults. Thus, the infant affords us the opportunity to observe the workings of the model presented in Figure 3, at an early stage of development, without the necessity of considering the whole system shown in Figure 2. In earlier days, many would have said we were dealing with the "prelinguistic" child (Kaplan and Kaplan, 1971). In some sense, however, this turns out not to be true; the work of Eimas (Eimas, Siqueland, Jusczyk, and Vigorito, 1971; Moffitt (1971), Morse, (1972), Eimas, (1974, 1975), and others, has shown that infants are quite sensitive to speech distinctions. They can discriminate phonetic distinctions such as those between /ba-/pa/ and /ba-/da/, and they cannot distinguish between members of the same phonemic category. These results are functionally parallel to those for

adults in categorical perception (see for example, Mattingly, Liberman, Syrdal, and Halwes, 1971). Moreover, young infants can distinguish between the initial liquid phonemes in /ra/ and /la/ (Eimas, 1975) better than adults in cultures where this distinction is not phonemic (Miyawaki, Strange, Verbrugge, Liberman, Jenkins, and Fujimura, 1975). It would appear that young infants come equipped with capacity to discriminate the relevant acoustic attributes that underlie almost all phonetic features.

The question remains, however, whether or not these infants are perceiving speech as speech (Stevens and Klatt, 1974). In terms of our information-processing model, where, for example, does such processing occur? A few years ago we would have said--unequivocally--that such discriminations must be accounted for on the basis of stage 2, the phonetic feature analyzers (see for example, Cutting and Eimas, 1975). Today, we are not so sure. In fact, most of the evidence for such discriminations points to stage 1, the auditory feature analyzers. We have previously claimed that the experiments on selective adaptation appear to be working at a stage prior to phonetic analysis; given that position, we suggest that the infant studies indicate perception at the same stage. These results indicate not so much that young infants perceive speech, as they indicate that they can perceive speech-relevant dimensions of an acoustic signal (see Jusczyk, Rosner, Cutting, Foard, and Smith, in press) that they will later apply to the process of speech perception. As Roger Brown (1973:37) might note, such perception in infants is "only linguistic by courtesy of its continuity with a system which in fully elaborated form is indeed..." speech perception.

An important aspect of infant research on speech perception that we see missing at this point is the performance by the infant, of many-to-one and one-to-many mapping of acoustic features onto phonetic features. Fodor, Garrett, and Brill (1975) have taken a small step in this direction using older infants: they found that four- and five-month old infants, like adults, perceived phonemic identities in the syllables /pi/ and /pu/ as different from the syllable /ka/, yet in all three syllables the voiceless stop consonant is cued by similar acoustic information (Schatz, 1954). Until such evidence for one-cue-to-many-phoneme mappings can be assembled more fully, along with corresponding many-to-one results, we regard the infant data as indicating speech-relevant perception, rather than speech perception.

#### Experimental Evidence with Animals

Further support for the allocation of sophisticated analyses to acoustic-feature processing stems from some recent studies with rhesus monkeys and chinchillas. Their discriminations of synthetic speech syllables differing in place of articulation (Morse and Snowden, 1975) and identifications along VOT (Kuhl and J. D. Miller, 1975) look suspiciously like categorical perception. While these results may be subject to range effects (see Parducci, 1974), and while some speech dimensions may not be perceived categorically by animals (Sinnott, 1974), it is clear that the once-firm base of empirical data thought to be indicative of phonetic perception may actually be a data base supporting the existence of a sophisticated auditory-property-analysis stage of processing (see Cutting and

Rosner, 1974; Cutting, Rosner and Foard, in press; Cutting, in press; J. D. Miller, Wier, Pastore, Kelly, and Dooling, in press). These animals, and perhaps the human infants as well, do not perceive speech as language, but as a multidimensional complex of acoustic events.

#### Experimental Evidence for Neurological Locus

Mapping the stages of an information-processing model onto neurological structure is not an easy task. While the clinician, in particular, needs to know about such facts that we are able to compile, we should all remember that fractionation of the speech/language system with regard to clinical populations can be a hindrance rather than an aid (Jenkins, Jimenez-Pabon, Shaw, and Sefer, 1975).

In general, the great majority of linguistic processes appear to be associated with the left cerebral hemisphere of the human brain (Geschwind, 1970). Whereas lexical, semantic, and syntactic operations may be best performed by this hemisphere, the right hemisphere also appears to play an important role in the perception of phonological cues such as stress (Blumstein and Goodglass, 1972) and intonation (Blumstein and Cooper, 1974). Thus, we should not rule out the dynamic role of both hemispheres in the language process, especially in view of the therapeutic value of exercising right-hemisphere functions on the recovery of language abilities after stroke (Albert, Sparks, and Helm, 1973; Keith and Aronson, 1975).

Phonetic level processing may occur, in part, in the right hemisphere as well as in the left hemisphere. Considering the stages of the recognition device shown in Figure 3, there is, first of all, no reason to assume that preliminary auditory storage is not perfectly bilateral, with equipotential left- and right-ear components. Most aspects of acoustic feature analysis (stage 1) may also be bilateral; but certain aspects may be best processed by the left hemisphere, such as the analysis of rapid frequency changes (Halperin, Nachshon, and Carmon, 1973; Cutting, 1974) and the processing of rapid amplitude modulations in the acoustic signal.<sup>1</sup> The remaining stages may reside entirely in the left hemisphere, but it is only the stage of phonetic feature combination (stage 4) for which this seems a logical necessity. Since combination can be seen as a "blending" of phonetic features, be they from two dichotic inputs or one binaural input, a single mixing device is needed. While this mixing could be duplicated in both hemispheres, it seems unlikely given the nature of the auditory pathways to each hemisphere (Milner, Taylor, and Sperry, 1968). If this device were in the right hemisphere it would be removed from many of the other aspects of language. Economy of design, then, would warrant placing it within the left hemisphere. Data supporting this allocation stem from electrophysiological (Wood, Goff, and Day, 1971; Wood, 1975) and dichotic-listening analyses (Studdert-Kennedy et al, 1972).

---

<sup>1</sup>Mark Blechner (1976): personal communication.



Short-term memory (STM) is certainly bilateral, but different forms of it may be hemispherically specialized. Verbal forms of STM information, for example, appear to occur in the left hemisphere, and spatial imagery forms of STM information appear to occur in the right hemisphere (Seamon and Gazzaniga, 1973). This should not surprise us. It is only a reflection of what the separate hemispheres appear to do best (Kimura, 1967).

#### SECTION 5: SOME CONTROVERSIES IN SPEECH PERCEPTION

The basic issues in speech perception today are nearly the same as they were 25 years ago--the apparent lack of invariance in the acoustic signal, and the related problems of segmentation and appropriate units for analysis. As for the issue of appropriate units for analysis, we can quickly point to Figures 2 and 3 and say that we deem all units appropriate--auditory features, phonetic features, phonemes, phonological features, morphemes, words, clauses, sentences, paragraphs, and so on. All are important in speech perception and all are used; the great body of literature in experimental psycholinguistics supports this view. Awareness (or reversing Polanyi's metaphor "linguistic opacity") of some of the speech units, however, appears to develop in the child only after language acquisition is well advanced (I. Liberman, Shankweiler, Fischer, and Carter, 1974; Mattingly, 1972), and delays onset of reading readiness. We will discuss this point further in Section 6.

Lack of invariance and the problem of segmentation are not as easily dealt with in our model. The acoustic representation of speech is a great accomplishment in parallel transmission of information (see Figure 5). Here, the words SANTA CLAUS are spectrographically displayed and parsed roughly according to acoustic segments (a). These acoustic segments are then mapped onto phonemic segments (b), and the amount of parallel transmission can be seen. Segment 8, for example, appears to carry information about four phonemes, /ntak/ in only 50 msec; the phone/n/ can be seen smeared across acoustic segments 3 through 8, a duration of roughly 200 msec. Thus, the acoustic shapes of each phoneme are folded into one another to a very great degree, confounding the problem of invariance in the acoustic display.

Not only is the acoustic shape of one phoneme contingent on its immediate neighbors, but also on phonemes that may be three or more phonemes away. For example, the /s/ in STREET is quite different from the /s/ in STROBE because, in anticipation of the different vowels /i/ and /o/, the lips are unrounded in the second. Such coarticulation can be so extreme that, as mentioned earlier, identical acoustic events can cue different phonemes in certain contexts, and quite different events can cue the same phoneme in others (Liberman, Cooper, Shankweiler, Studdert-Kennedy, 1967).

Some have viewed this position as overstated, and they have a point, especially in view of the time-window perceptual-moment hypothesis discussed previously. Burst cues, for example, are often only 5 to 10 msec in duration, well below auditory-resolution thresholds, and certainly must integrate into following acoustic segments. Cole and Scott (1974a; 1974b) have claimed that invariant cues exist even for the most highly variable

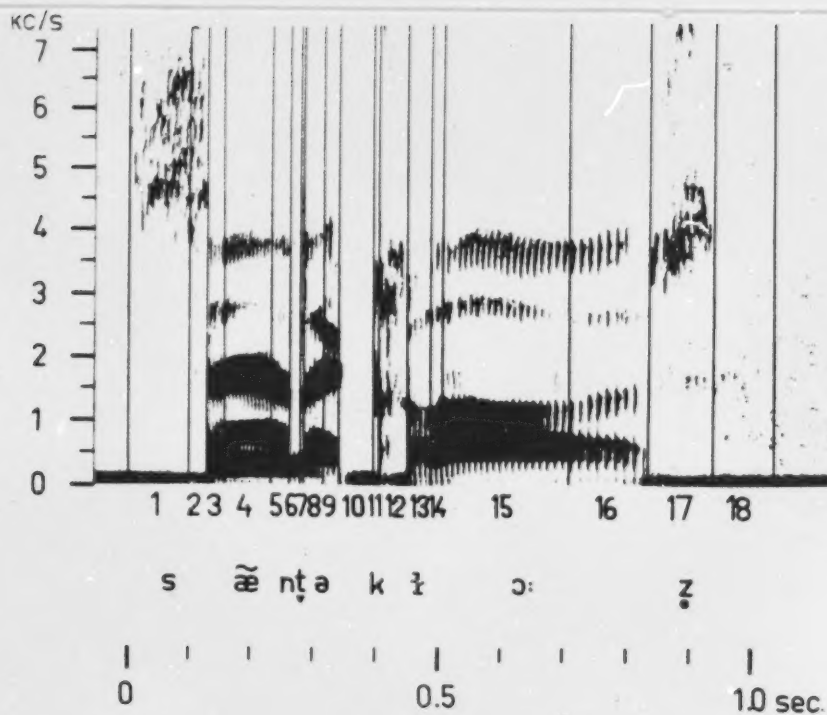
phonemes, the stop consonants. This is an epistemological position that Malcolm (1971), for one, would like. If true, it makes possible a simple, direct approach to speech perception without circuitous interconnections of stages and without need of overt contributions of the human perceiver. Recent evidence, however, demonstrates that while there may be extremely brief invariant bursts for stops, they are often largely unusable in real-time speech perception.<sup>2</sup> Thus, in our opinion, pure template-matching accounts of the speech process are not entirely adequate, despite their attractiveness.

Without great amounts of invariance, then, speech perception seems to be an impossible endeavor. Although some significant aspects of the signal are invariant--for example, certain aspects of some stressed vowels, of fricatives, and of nasals--many simply are not. How, then, is speech perceived? Two allied views that suggest a way to cut through this Gordian knot are the "motor theory" (Liberman, et al, 1967; Studdert-Kennedy, Shankweiler and Schulman, 1970; Cooper, 1972) and analysis by synthesis (Stevens, 1960, 1972; Stevens and House, 1972). Both are dynamic views of speech perception and both can be placed easily within the information-processing scheme of Figures 2 and 3. In the motor theory, the invariance problem is thought to be resolved at the neuromotor level, while in analysis by synthesis, it is resolved at the neuroacoustic level. Both accounts, when applied to our model, would add the assumption that many-to-one and one-to-many mappings of auditory properties onto phonetic features are done with tacit knowledge of articulation (the motor theory) or its acoustic consequences (analysis by synthesis). Kuhn (1975) has suggested a way in which the latter might be done. A portion of the supralaryngeal vocal tract, the front cavity, may provide higher-order invariant information (see Hochberg, 1974) of a kind different from that suggested by Cole and Scott (1974a). The front cavity may allow straightforward computation of place of articulation and vowel information through selective perception of the second formant, or a weighted combination of the second and third formants. The only knowledge necessary for perception of place of articulation and vowel nucleus, then, is the general size of the vocal tract. The importance of the front cavity resonance to perception has yet to be explored fully, but it may bring us closer to understanding the transformation from auditory property to phonetic feature. We believe that it is knowledge of this transformation that is crucial to understanding of speech perception.

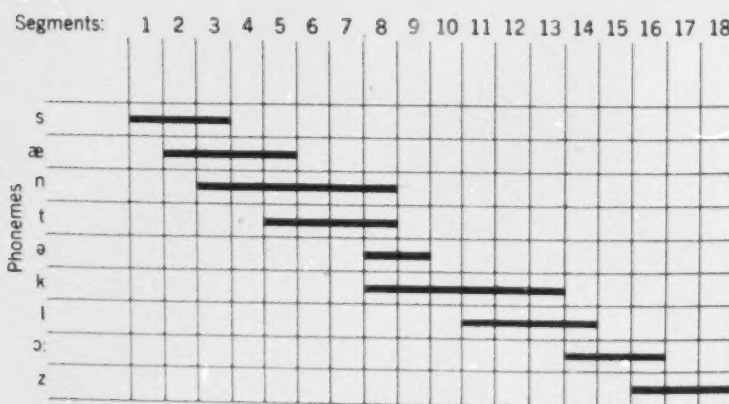
Segmentation, according to motor theory and analysis-by-synthesis views with regard to articulation, is also accomplished at some level. We parse the incoming speech stream according to the way speech is produced. Thus, all the stages within the recognition device would appear to have access to knowledge about speech gestures or their resultant effects on acoustic shape.

---

<sup>2</sup>M. Dorman, M. Studdert-Kennedy, and L. Raphael (1976): personal communication.



(a)



(b)

Figure 5: Perceptually, speech sounds seem to follow one another like a train of independent speech segments. Acoustically, however, there is considerable overlap. A spectrogram of the words SANTA Claus (a), where vertical lines mark acoustically different segments, and the assignment of phonemes to those segments (b) (from Fant and Lindblom, 1961).

As a final point, it should be noted that knowledge of articulation or articulatory consequences need not be based on the ability to articulate. Thus, Lenneberg's (1962) case study of a speechless but linguistically sophisticated child is not counterproof of role of articulation in speech perception. We reemphasize that most of the infant speech perception results and all of the "speech" perception results with animals appear to indicate speech-relevant auditory processing rather than phonetic processing. It is simply not important, then, that the infant's vocal tract and the animal's vocal tract are dissimilar from that of adult humans (Liberman, Crelin, and Klatt, 1972). Current criticisms of articulation-mediated speech perception that cite such findings (see, for example, Palermo, 1975) are not decisive.

This view assumes that underlying categorical perceptions are auditorily based, rather than phonetically based decision processes. That is, categorical perception may be accounted for at Stage 2 of our model, rather than at Stage 3, as assumed in the past. Recent evidence with complex acoustic signals not carrying linguistic information appears to support a psychophysical basis for categorical perception (Cutting and Rosner, 1974; J. D. Miller, Wier, Pastore, Kelly, and Dooling, in press; Pisoni, 1976). Are the outputs from the separate auditory property detectors always discrete and categorical? At this point we simply do not know. It may be that more "continuous" perception, such as that found for vowels (Pisoni, 1973), is distinguished from categorical perception solely by the interaction of the roles of echoic memory and of auditory property analysis.

#### SECTION 6: IMPLICATIONS FOR THE SCHOOL AND CLINIC

##### Auditory Processes

In considering the possible implications of this model for school and clinic situations, we will limit ourselves to phonetic perception as shown in Figure 3. Whereas our model is based on the normal speaker/listener, it should apply to certain special populations if we allow the modification of dynamic relationships within the model. Before beginning, however, a few caveats are in order. First, we would be surprised to find a single person whose language deficit could be attributed to the malfunction of a single stage in this model. The model itself would predict that, under a long lasting deficit, anomalies would develop that run through the entire system. Second, this model may be discarded in its entirety when applied to the person who has been profoundly deaf since birth. If this individual substituted sign language for speech, the phonetic aspect of the model would have no relevance. Recent developments in the linguistic analysis of sign language have shown it to be quite different from speech (Bellugi and Fischer, 1972; Bellugi, Klima, and Siple, 1974/1975; Bellugi and Klima, 1975; Frishberg, 1975; Friedman, 1975; Lane, Boyes-Braem, and Bellugi, 1976).

The first important stage to be considered is the initial registration of the signal, or preliminary auditory analysis. Put more simply, we must consider certain aspects of hearing and deafness. As Danaher, Osberger, and Pickett (1973:439) have noted: "Most persons with sensorineural hearing loss have some impairment of speech discrimination ability," and generally the



discrimination of consonants is poorer than that of vowels. This difference could stem from several causes.

There are many kinds of sensorineural deafness. Persons can be classified as having "flat" fairly constant losses across the frequency spectrum, or "sloping" losses that increment substantially for each octave above 500 Hz (Danaher, Osberger and Pickett, 1973). Moreover, these pure-tone audiograms are not related in any simple way to speech-sound discrimination (Danaher and Pickett, 1975). Thus, even at this first stage of analysis we appear to be awash with complexity. Part of the resolution must be that we are not dealing with a single stage of perceptual analysis.

Danaher and Pickett (1975) note three types of masking that appear to reduce the discriminability of second-formant transitions. The first is a simple upward spread of masking of peripheral origin. That is, low frequency components of the speech signal mask higher frequency components. These individuals improve their discriminations when the signal level of the first formant is reduced by 10 dB, or when the lower frequency components of the signal are presented to one ear and the upper frequency components to the other. Presenting the first formant to one ear and the upper formants to the other, as Broadbent (1955) and others (Rand, 1974; Cutting, 1976) have done in nonclinical settings, does not alleviate upward spread of masking in all individuals. In this second group, dichotic split-formant presentation is of little help. A third type of masking, as Danaher and Pickett (1975) note, complicates the issue even further. It is a type of simple backward masking in which information in a steady-state portion of a vowel reduces the detectability of initial formant transitions.

Applying these three types of data to our model cannot be straightforward, especially since hard-of-hearing individuals differ in their clinical histories. It appears that--whereas decrements due to preliminary auditory analysis probably occur in all three types of masking--differential decrements at other stages may also be involved. We suspect that the difference between monaural and dichotic upward-spread of masking may be attributable to differences in ability to perform auditory feature analysis. In the monaural case, the decrement in performance may be attributable to preliminary auditory analysis, whereas in the case of dichotic-upward spread, the decrement may be attributable to that stage as well as to auditory feature analysis. If it is possible that the two groups of people, those who do and those who do not demonstrate dichotic masking, differ in the duration of their hearing loss, the latter group's auditory feature-analyzers may have become inoperative through longer-term disuse. We are speculating here, but the dichotic situation is an unusual one, not experienced outside the laboratory. It may be that these auditory feature-analyzers have not been stimulated for a long period of time and have "atrophied" (Eimas, 1975).

Hearing loss attributable to backward masking would appear to be of a different kind. Backward masking is a temporal phenomenon, whereas upward spread of masking is a frequency one. Temporal phenomena implicate memories and integrating time windows, and we suspect that part of this loss of information is attributable to limitations on coding information in sensory

information store, as well as auditory feature analysis.

This assumption appears to be bolstered by results of research on quite a different population. Tallal and Piercy (1973; 1974; 1975) have studied children diagnosed as developmental aphasics (see also Tallal, in press). These children have normal audiograms but have extreme difficulty perceiving temporally patterned auditory signals. In a series of careful studies, they have found that the developmental aphasic has much more difficulty discriminating consonants than vowels, and that cue duration appears to be the primary cause. Short vowels (43 msec) are more difficult to discriminate than long vowels (95 msec) when the target vowel was followed by a masking vowel in a situation with no target-mask interstimulus interval. This result is interesting since normal listeners often report no difficulty in identifying target vowels of even shorter duration in a similar situation (Dorman, Kewley-Port, Brady, and Turvey, in press).

The short and long vowel results from these children appear to implicate a deficit in sensory information store, and perhaps in auditory property analysis. As Tallal and Piercy point out, these children appear to have an auditory system whose early stages are more sluggish than the norm. Perhaps their integrating time windows are larger. Without remediation of this difficulty, deficits throughout the system, especially at the higher levels, are likely to persist, even spread, as Tallal (in press) suggests.

We propose that the deficits looked at thus far are not deficits of speech perception per se, but deficits primarily in auditory analysis. Clearly, these deficits have effects higher in the system, but we think that those result from anomalies at this lower level and not from upper level causes.

#### Phonetic Processes?

Evidence for phonetic impairment without auditory impairment is much more difficult to acquire. In fact, given the recent importance attributed to the stage of auditory feature analysis, there may be no such evidence at the present time: results from normal listeners that used to be considered as indicative of phonetic processing (see Wood, 1975:16) now appear to be indicative of auditory feature analysis (see Cutting et al., in press). Evidence for stage 2 may be contaminated by the operation of prior stages, or even subsequent ones. Let us consider, then, some of the problems of such contamination and the more general problems of applying experimental paradigms to field situations in the school and clinic.

The standard dichotic listening procedure used in the laboratory during the last fifteen years may not yield much useful information for workers in the school and clinic. The reasons for this assessment stem from several sources. First, it is difficult and sometimes unwise, to lump together persons with similar perceptual problems. Given the great complexity of the human organism, these individuals are likely to differ greatly even if they fit into the same diagnostic category. Witness the results of Danaher and Pickett (1975). Thus, an individual-oriented procedure would seem best;

however, this is the point at which the utility of the dichotic listening procedure and most other laboratory techniques become strained. In dichotic listening, a right-ear advantage may be attributable to a number of causes, including those associated with almost every stage shown in Figures 2 and 3. The existence of a right-ear advantage would tell the applied researcher that some part of the speech/language system is running properly. Even by using nonsense syllables as stimuli, the applied researcher may gain no more information than that (Dorman and Geffner, 1974; Dorman and Porter, 1975). The lack of a right-ear advantage may be indicative of malfunction in one stage, but even in the best of situations it does not tell the researcher which stage, and in the worst of situations it does not imply that any stage is malfunctioning. The existence of a right-ear advantage even among the normal population is only a probabilistic occurrence, and the lack of an ear advantage should not be seen as an abnormality (Shankweiler and Studdert-Kennedy, 1975). We feel strongly about this, if only for the reason that one of the authors consistently yields no ear advantage in these tasks. Moreover, left-ear advantages for linguistic material are not uncommon, and must be interpreted with caution even in the most extreme instances (see Fromkin, Krashen, Curtiss, Rigler, and Rigler, 1974).

Results from other paradigms may be more interpretable, but they may not indicate much about phonetic processing. The application of identification and discrimination paradigms associated with categorical perception, selective adaptation, and tasks requiring selective attention to different dimensions of the speech stimulus, may be valuable in terms of knowledge about general auditory processing and may serve as verification (or falsification) of certain aspects of our model.

#### Units of Speech Analysis and Reading

While all units of speech analysis are relevant to speech perception, not all of them are equally apparent, especially to the child. Conscious access to, or knowledge of, speech parsing routines is important in the acquisition of reading skill. In logographic writing, where words are the primary unit such as in Chinese script, and in the Japanese use of Kanji script, knowledge about words is imperative. In syllabary writing, where syllables are the primary unit, such as in the Japanese use of Kana script and Cherokee use of the script invented by Sequoia, knowledge of syllables is imperative. In alphabetic writing, where the phoneme is the principal unit as in English, knowledge of phonemes is imperative.

To the young child, syllables and phonemes are not equally amenable units of analysis. I. Liberman, Shankweiler, Fischer, and Carter (1974) examined the ability of nursery schoolers, kindergarteners, and first graders to tap out the number of syllables and phonemes in common words. For each child the number of trials needed to reach a criterion of six consecutive correctly tapped trials without demonstration by the experimenter, was measured. Ability to segment by syllable was shown by half of the four-year olds, but none of them could segment by phoneme. By age six, 90 percent of the children could segment by syllable, but only 70 percent by phoneme. From these results, Liberman et al suggested that one of the reasons that children



under the age of five or six are not ready to learn to read is because they are not yet consciously aware of the units on which written English is based.

Another important aspect of reading readiness, as Conrad (1972) notes, is the ability to verbally code information into short-term memory. I. Liberman, Shankweiler, Liberman, Fowler, and Fischer (in press) applied this notion to good and poor beginning readers. They suspected that good beginning readers would be those using their newly acquired skill of parsing language into phonemes, and developing the appropriate coding in short-term memory based on phonemic structure. Poor readers, on the other hand, might still be baffled by phonemes, and therefore unable to use the phoneme code. These two groups of beginning readers, equated for intelligence, were presented with two types of consonant strings to view. One type of string consisted of consonants with rhyming names (b, c, d, g, p, t, v, z), and the other with nonrhyming names (h, k, l, q, r, s, w, y). Subjects were then tested for conditions of immediate recall and delayed recall. In general, the good readers made fewer errors than the poor readers on both rhyming and nonrhyming consonants, although the rhyming consonants were more difficult for all. More importantly, the advantage shown by the good readers disappeared for the rhyming consonants in the delayed recall condition, but did not disappear for the nonrhyming consonants. The implication is that the decrement in the delayed condition shown by the good readers, but not by the poor readers, is attributable to the use of phonetic codes in short-term memory by good readers and the possible use of some other code, perhaps a visual one (Conrad, 1972), by the poor readers.

The implication of these results for the school and clinic is, that for the beginning reader of English, reading is a derivative of speech perception. Acquisition of reading skill depends, in part, on mastery of certain linguistic skills and on the awareness by the youngster of the proper units of analysis. In order to learn to read an alphabetically-written language, the availability of a phonetically organized short-term memory may not be enough. Added to that memory must be phoneme parsing skills. Reading instruction and remediation programs should note: awareness of speech segments should be an early goal of instruction.

#### SECTION 7: SUMMARIZING REMARKS

Our goal has been to present one point of view about the perceptual analysis of speech sounds. It is a process that takes time, consists of many stages, uses a number of memory stores, is limited in certain ways, and uses all levels of the speech/language system, including phonetic, phonological, lexical, syntactic, and semantic components. We have presented evidence that supports our model of speech perception from research on adult and infant humans, as well as on animals. This evidence, of course, does not confirm our model while disconfirming alternative views, but we think that the evidence plus important logical considerations make our view most plausible. Speech perception and speech production appear to be inextricably intertwined--on grounds of parsimony as Lashley suggested, and on grounds that there is too complex a mapping from auditory features to phonetic features, and from phonetic features to phonological features--for any



alternative method of speech perception to be feasible. Finally, we have taken some of the general views and findings of speech perception research and tried to apply them to the school and clinic. Certain experiments on aphasic and hard-of-hearing populations may help support our view of speech perception. Other findings of speech research, especially those connected with reading, may help the applied researcher in the school and clinic.

#### REFERENCES

- Albert, M. L., R. W. Sparks, and N. A. Helm. (1973) Melodic intonation therapy for aphasia. Arch. Neurol. 29, 130-131.
- Allport, D. A. (1968) Phenomenal simultaneity and the perceptual moment hypothesis. Brit. J. Exp. Psychol. 59, 395-406.
- Barclay, J. R. (1975) Noncategorical perception of a voiced stop: A replication. Percept. Psychophys. 11, 269-273.
- Bellugi, U. and S. Fischer. (1972) A comparison of sign language and spoken language. Cognition 1, 173-200.
- Bellugi, U. and E. S. Klima. (1975) Aspects of sign language structure. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press), pp. 171-203.
- Bellugi, U., E. S. Klima, and P. Siple. (1974/1975) Remembering in signs. Cognition 3, 93-125.
- Binnie, C. A., A. A. Montgomery, and P. L. Jackson. (1974) Auditory and visual contributions to the perception of selected English consonants. J. Speech Hearing Res. 17, 619-630.
- Bjork, R. A. (1975) Short-term storage: The ordered output of a central processor. In Cognitive Theory, vol. 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.), pp. 151-172.
- Blakemore, C. and F. W. Campbell. (1969) On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. J. Physiol. 203, 237-260.
- Blumstein, S. and W. E. Cooper. (1974) Hemispheric processing of intonation contours. Cortex 10, 146-158.
- Blumstein, S. and H. Goodglass. (1972) The perception of stress as a semantic cue in aphasia. J. Speech Hearing Res. 15, 800-806.
- Bransford, J. and K. E. Nitsch. (in press) How can we come to understand things that we did not previously understand? In "Implications of Basic Speech and Language Research for the School and Clinic," conference proceedings, ed. by J. F. Kavanagh and W. Strange. (Cambridge, Mass.: MIT Press).
- Bresnan, J. (1976) Toward a realistic model of transformational grammar. (Paper presented at the conference on New Approaches to a Realistic Model of Language, MIT, Cambridge, Mass.), March 9.
- Broadbent, D. E. (1955) A note on binaural fusion. Quart. J. Exp. Psychol. 7, 46-47.
- Broadbent, D. E. (1958) Perception and Communication. (London: Pergamon Press).
- Broadbent, D. E. (1965) Information processing in the nervous system. Science 150, 457-462.
- Broadbent, D. E. (1971) Decision and Stress. New York: Academic Press).

- Broadbent, D. E. (1975) The magic number seven after fifteen years. In Studies in Long Term Memory, ed. by A. Kennedy and A. Wilkes. (London: Wiley), pp. 3-18.
- Brown, R. (1973) A First Language. (Cambridge, Mass: Harvard University Press).
- Brown, R. and D. McNeill. (1966) The "tip of the tongue" phenomenon. J. Verbal Learn. Verbal Behav. 5, 325-337.
- Chomsky, N. (1964) Current issues in linguistic theory. In The Structure of Language, ed. by J. A. Fodor and J. J. Katz. (Englewood Cliffs, N.J.: Prentice-Hall), pp. 50-118.
- Cole, R. A. (1973) Listening for mispronunciations: A measure of what we hear during speech. Percept. Psychophys. 13, 153-156.
- Cole, R. A. and B. Scott. (1974a) The phantom in the phoneme: Invariant cues for stop consonants. Percept. Psychophys. 15, 101-107.
- Cole, R. A. and B. Scott. (1974b) Toward a theory of speech perception. Psychol. Rev. 81, 348-374.
- Collins, A. M. and M. R. Quillian. (1969) Retrieval time from semantic memory. J. Verb. Learn. Verbal. Behav. 8, 240-248.
- Conrad, R. (1972) Speech and reading. In Language by Ear and by Eye, ed. by J. F. Kavanagh, and I. G. Mattingly. (Cambridge, Mass.: MIT Press), pp. 205-240.
- Cooper, F. S. (1972) How is language conveyed by speech? In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press), pp. 25-46.
- Cooper, W. E. (1975) Selective adaptation to speech. In Cognitive Theory, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni, vol. 1. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.), pp. 23-54.
- Cooper, W. E. and S. E. Blumstein. (1974) A "labial" feature analyzer in speech perception. Percept. Psychophys. 15, 591-600.
- Crowder, R. B. (1971) The sound of vowels and consonants in immediate memory. J. Verbal Learn. Verbal Behav. 10, 587-596.
- Crowder, R. B. (1972) Visual and auditory memory. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press), pp. 251-276.
- Crowder, R. B. (in press) Language and memory. In "Implications of Basic Speech and Language Research for the School and Clinic" conference proceedings, ed. by J. F. Kavanagh and W. Strange. (Cambridge, Mass.: MIT Press).
- Crowder, R. G. and J. Morton. (1969) Precategorical acoustic storage (PAS). Percept. Psychophys. 5, 365-373.
- Cutting, J. E. (1974) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys. 16, 601-612.
- Cutting, J. E. (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychol. Rev. 83, 114-140.
- Cutting, J. E. (in press) The magical number two and the natural categories of speech and music. In Tutorial Essays in Psychology, ed. by N. S. Sutherland. (Hillsdale, N.J.: Lawrence Erlbaum Associates).
- Cutting, J. E. and P. D. Eimas. (1975) Phonetic feature analyzers and the processing of speech in infants. In The Role of Speech in Language,

- ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press), pp. 127-148.
- Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.
- Cutting, J. E., B. S. Rosner, and C. F. Foard. (in press) Perceptual categories for musiclike sounds: Implications for theories of speech perception. Quart. J. Exper. Psychol. 28.
- Danaher, E. M., M. J. Osberger, and J. M. Pickett. (1973) Discrimination of formant frequency transitions in synthetic vowels. J. Speech Hearing Res. 16, 439-451.
- Danaher, E. M. and J. M. Pickett. (1975) Some masking effects produced by low-frequency vowel formants in persons with sensorineural loss. J. Speech Hearing Res. 18, 261-271.
- Darwin, C. J. and A. D. Baddeley. (1974) Acoustic memory and the perception of speech. Cog. Psychol. 6, 41-60.
- Day, R. S. and C. C. Wood. (1972) Interactions between linguistic and nonlinguistic processing. J. Acoust. Soc. Am. 51, 79(A).
- DeCordomoy, G. (1668) A Philosophical Discourse Concerning Speech. (London: J. Martin).
- Dorman, M. F. and D. S. Geffner. (1974) Hemispheric specialization for speech perception in six-year-old black and white children from low and middle socio-economic classes. Cortex 10, 171-176.
- Dorman, M. F., D. Kewley-Port, S. Brady, and M. T. Turvey. (in press) Vowel recognition: Inferences from studies of forward and backward masking. Quart. J. Exper. Psychol.
- Dorman, M. F. and R. J. Porter. (1975) Hemispheric lateralization for speech perception in stutterers. Cortex 11, 181-185.
- Eimas, P. D. (1974) Auditory and linguistic processing of cues for place of articulation by infants. Percept. Psychophys. 16, 513-521.
- Eimas, P. D. (1975) Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. Percept. Psychophys. 18, 341-347.
- Eimas, P. D., W. E. Cooper, and J. D. Corbit. (1973) Some properties of linguistic feature detectors. Percept. Psychophys. 13, 206-217.
- Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.
- Eimas, P. D., E. R. Siqueland, P. W. Jusczyk, and J. M. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.
- Erber, N. P. (1969) Interaction of audition and vision in the recognition of oral speech stimuli. J. Speech Hearing Res. 12, 423-425.
- Erber, N. P. (1975) Auditory-visual perception of speech. J. Speech Hearing Dis. 40, 481-491.
- Fant, G. and B. Lindblom. (1961) Studies of minimal speech and sound units. Quarterly Progress Report, (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) 2, 1-11.
- Fodor, J. A., T. G. Bever, and M. F. Garrett. (1974) The Psychology of Language. (New York: McGraw-Hill).
- Fodor, J. A., M. F. Garrett, and S. L. Brill. (1975) Pi ka pu: The perception of speech sounds by prelinguistic infants. Percept. Psychophys. 18, 74-78.
- Foulke, E. and T. G. Sticht. (1969) Review of research on the intelligi-

- bility and comprehension of accelerated speech. Psychol. Bull. 72, 50-62.
- Friedman, L. A. (1975) Space, time, and person reference in American sign language. Lang. 51, 940-961.
- Frishberg, N. (1975) Arbitrariness and iconicity: Historical change in American sign language. Lang. 51, 696-719.
- Fromkin, V. A. (1971) The non-anomalous nature of anomalous utterances. Lang. 47, 27-52.
- Fromkin, V. A., S. Krashen, S. Curtiss, D. Rigler, and M. Rigler. (1974) The development of language in Genie: A case of language acquisition beyond the "critical period." Brain Lang. 1, 81-107.
- Geschwind, N. (1970) The organization of language and the brain. Science 170, 940-944.
- Gerstman, L. J. (1957) Perceptual dimensions for the friction portion of certain speech sounds. (PhD dissertation, New York University).
- Haber, R. N. (1969) Introduction. In Information-processing Approaches to Visual Perception, ed. by R. N. Haber. (New York: Holt, Rinehart, Winston), pp. 1-15.
- Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift in ear superiority in dichotic listening to temporal pattern nonverbal stimuli. J. Acoust. Soc. Am. 53, 46-50.
- Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. (PhD dissertation, University of Minnesota). Dissertation Abstracts International, 1970, 31, 1565b. (Univ. Microfilms 70-15, 736).
- Hirsh, I. J. (1974) Temporal order and auditory perception. In Sensation and measurement: Papers in Honor of S. S. Stevens, ed. by H. R. Moskowitz, B. Scharf, and J. C. Stevens. (Boston: D. Riedel), pp. 51-258.
- Hochberg, J. (1974) Higher-order stimuli and inter-response coupling in the perception of the visual world. In Perception: Essays in Honor of James J. Gibson, ed. by R. B. MacLeod and H. L. Pick. (Ithaca, N.Y.: Cornell University Press), pp. 17-39.
- Howes, D. (1957) On the relation between the intelligibility and frequency of occurrence of English words. J. Acoust. Soc. Am. 29, 296-305.
- Hughes, G. W. and M. Halle. (1956) Spectral properties of fricative consonants. J. Acoust. Soc. Am. 28, 303-310.
- Jakobson, R., G. Fant, and M. Halle. (1951) Preliminaries to Speech Analysis, 1963, 3rd printing. (Cambridge, Mass.: MIT Press).
- Jenkins, J. J., E. Jimenez-Pabon, R. E. Shaw, and J. W. Sefer. (1975) Schuell's Aphasia in Adults. (New York: Harper & Row).
- Juszyk, P. W., B. S. Rosner, J. E. Cutting, C. F. Foard, and L. Smith. (in press) Categorical perception of nonlinguistic sounds in the two month old infant. Percept. Psychophys.
- Kaplan, E. L. and G. Kaplan. (1971) The prelinguistic child. In Human Development and Cognitive Processes, ed. by J. Eliot. (New York: Holt, Rinehart, and Winston), pp. 358-381.
- Keith, R. L. and A. E. Aronson. (1975) Singing as therapy for apraxia of speech and aphasia: report of a case. Brain Lang. 2, 483-488.
- Kimura, D. (1967) Dual functional asymmetry of brain in visual perception. Neuropsychologia 4, 275-285.



- Kuhl, P. A. and J. D. Miller. (1975) Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. Science 190, 69-72.
- Kuhn, G. M. (1975) On the front cavity resonance and its possible role in speech perception. J. Acoust. Soc. Am. 58, 428-433.
- Lane, H., P. Boyes-Braem, and U. Bellugi. (1976) Preliminaries to a distinctive feature analysis of handshapes in American Sign Language. Cog. Psychol. 8, 263-289.
- Lashley, K. S. (1951) The problem of serial order in behavior. In Cerebral Mechanisms in Behavior, ed. by L. A. Jeffress. (New York: Wiley), pp. 112-136.
- Lenneberg, E. H. (1962) Understanding language without ability to speak: a case report. J. Abnormal Social Psychol. 65, 419-425.
- Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.
- Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
- Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1952) The role of selected stimulus variables in the perception of the unvoiced stop consonants. Am. J. Psychol. 65, 497-516.
- Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops. Lang. Speech 1, 153-167.
- Liberman, A. M., K. S. Harris, P. D. Eimas, L. Lisker, and J. Bastian. (1961) An effect of learning speech perception: The discrimination of durations of silence with and without phonemic significance. Lang. Speech 4, 175-195.
- Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.
- Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (New York: Winston), pp. 307-334.
- Liberman, I. Y., D. P. Shankweiler, F. W. Fischer, and B. Carter. (1974) Reading and the awareness of linguistic segments. J. Exp. Child Psychol. 18, 201-212.
- Liberman, I. Y., D. P. Shankweiler, A. M. Liberman, C. Fowler, and F. W. Fischer. (in press) Phonetic segmentation and recoding in the beginning reader. In Reading: Theory and Practice, ed. by A. S. Reber and D. Scarborough. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Licklider, J. C. R. (1952) On the process of speech perception. J. Acoust. Soc. Am. 24, 590-594.
- Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn human, Neanderthal man, and the chimpanzee. Am. Anthropol. 74, 287-307.
- Lisker, L. (1975) Is is VOT or a first-formant transition detector? J. Acoust. Soc. Am. 57, 1547-1551.
- Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384-422.
- Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice

- onset time in English stops. Lang. Speech 10, 1-28.
- MacKay, D. G. (1970) Spoonerisms: The anatomy of errors in the serial order of speech. Neuropsychologia 8, 323-350.
- Malcolm, N. (1971) The myth of cognitive processes and structures. In Cognitive Development and Epistemology, ed. by T. Mischel. (New York: Academic Press), pp. 385-392.
- Marslen-Wilson, W. D. (1975) Sentence perception as an interactive parallel process. Science 189, 226-228.
- Massaro, D. W. (1975) Language and information processing. In Understanding Language, ed. by D. W. Massaro. (New York: Academic Press), pp. 3-28.
- Massaro, D. W., and M. M. Cohen. (1975) Preperceptual auditory storage in speech perception. In Structure and Process in Speech Perception, ed. by A. Cohen and S. G. Nooteboom. (Heidelberg: Springer Verlag), pp. 226-243.
- Mattingly, I. G. (1972) Reading, the linguistic process, and linguistic awareness. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press), pp. 133-148.
- Mattingly, I. G., A. M. Liberman, A. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.
- Miller, G. A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81-96.
- Miller, G. A. (1962) Decision units in the perception of speech. IRE Transactions on Information Theory IT8, 81-83.
- Miller, G. A. (1972) English verbs of motion: A case study in semantics and lexical memory. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (New York: Winston), pp. 335-372.
- Miller, G. A. (in press) Lexical meaning. In "Implications of Basic Speech and Language Research for the School and Clinic" conference proceedings, ed. by J. F. Kavanagh and W. Strange. (Cambridge, Mass.: MIT Press).
- Miller, G. A., G. A. Heise, and W. Lichten. (1951) The intelligibility of speech as a function of the context of the test materials. J. Acoust. Soc. Am. 41, 329-335.
- Miller, G. A. and S. Isard. (1963) Some perceptual consequences of linguistic rules. J. Verbal Learn. Verbal Behav. 2, 217-228.
- Miller, G. A. and P. Nicely. (1955) An analysis of some perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27, 338-352.
- Miller, J. D., C. C. Wier, R. E. Pastore, W. J. Kelly, and R. J. Dooling. (in press) Discrimination and labeling of noise-buzz sequences with varying noise-lead times: an example of categorical perception. J. Acoust. Soc. Am.
- Miller, J. L. (1975) Properties of feature detectors for speech: evidence from the effects of selective adaptation on dichotic listening. Percept. Psychophys. 18, 389-397.
- Miller, J. L. and P. D. Eimas. (1976) Studies on the selective tuning of feature detectors for speech. J. Phonetics 4, 119-127.
- Milner, B., L. Taylor, and R. W. Sperry. (1968) Lateralized suppression

- of dichotically presented digits after commissural section in man. Science 161, 184-185.
- Miyawaki, K., W. Strange, R. Verbrugge, A. M. Liberman, J. J. Jenkins, and O. Fujimura. (1975) An effect of linguistic experience: the discrimination of [r] and [l] by native speakers of Japanese and English. Percept. Psychophys. 18, 331-340.
- Moffit, A. R. (1971) Consonant cue perception by twenty- to twenty-four-week-old infants. Child Develop. 42, 717-731.
- Morse, P. A. (1972) The discrimination of speech and nonspeech stimuli in early infancy. J. Exper. Child Psychol. 14, 477-492.
- Morse, P. A. and C. T. Snowden. (1975) An investigation of categorical speech discrimination by rhesus monkeys. Percept. Psychophys. 17, 9-16.
- Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).
- Norman, D. A. (1972) The role of memory in understanding language. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press), pp. 277-288.
- O'Neill, J. J. (1954) Contribution to the visual components of oral symbols to speech comprehension. J. Speech Hearing Dis. 19, 429-439.
- Orr, D. B., H. L. Friedman, and J. C. Williams. (1965) Trainability of listening comprehension of speeded discourse. J. Educ. Psychol. 56, 148-156.
- Palermo, D. S. (1975) Developmental aspects of speech perception: problems for a motor theory. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press), pp. 149-154.
- Parducci, A. (1974) Contextual effects: a range-frequency analysis. In Handbook of Perception, ed. by E. C. Carterette and M. P. Friedman, vol. 2. (New York: Academic Press), pp. 127-141.
- Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.
- Pisoni, D. B. (1975a) Dichotic listening and processing phonetic features. In Cognitive Theory, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni, vol. 1. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.), pp. 79-102.
- Pisoni, D. B. (1975b) Information processing and speech perception. In Speech Communication, ed. by G. Fant, vol. 3. (New York: John Wiley & Sons), pp. 331-337.
- Pisoni, D. B. (1976) Identification and discrimination of relative onset time of two-component tones: implications for voicing perception in stops. Quarterly Progress Report (Cambridge, Mass.: MIT Research Laboratory of Electronics).
- Pisoni, D. B. (in press a) Mechanisms of auditory discrimination and coding of linguistic information. In Second Auditory Processing and Learning Disabilities Symposium, ed. by J. V. Irwin. (Memphis, Tenn.: Memphis State University).
- Pisoni, D. B. (in press b) Speech perception. In Handbook of Learning and Cognitive Processes, ed. by W. K. Estes, vol 5. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.).
- Pisoni, D. B. and J. R. Sawusch. (1975) Some stages of processing in

- speech perception. In Structure and Process in Speech Perception, ed. by A. Cohen and S. Neebboom. (Heidelberg: Springer-Verlag), pp. 16-34.
- Pisoni, D. B. and J. B. Tash. (1975) Auditory property detectors and processing places features in stop consonants. Percept. Psychophys. 18, 401-408.
- Polanyi, M. (1964) Personal Knowledge. (New York: Harper & Row).
- Pollack, I. and J. M. Pickett. (1963) The intelligibility of excerpts from conversation. Lang. Speech 6, 165-171.
- Pollack, I. and J. M. Pickett. (1964) Intelligibility of excerpts from fluent speech: auditory vs. structural context. J. Verbal Learn. Verbal Behav. 3, 79-84.
- Rand, T. C. (1974) Dichotic release from masking for speech. J. Acoust. Soc. Am. 55, 678-680.
- Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. J. Acoust. Soc. Am. 51, 1296-1303.
- Savin, H. B. (1963) Word-frequency effect and errors in the perception of speech. J. Acoust. Soc. Am. 35, 200-206.
- Sawusch, J. R. (1976) The structure and flow of information in speech perception. (PhD. dissertation, Indiana University).
- Schatz, C. (1954) The role of context in the perception of stop. Language 30, 47-56.
- Seamon, J. G. and M. S. Gazzaniga. (1973) Coding strategies and cerebral laterality effects. Cog. Psychol. 5, 249-256.
- Shankweiler, D. P. and M. Studdert-Kennedy. (1975) A continuum of lateralization for speech perception. Brain Lang. 2, 212-225.
- Shiffrin, R. M. (1975) Short-term store: the basis for a memory system. In Cognitive Theory, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni, vol. 1. (Hillsdale, N.J.: Lawrence Erlbaum Assoc.), pp. 193-218.
- Shiffrin, R. M., D. B. Pisoni, and K. Casteneda-Mendez. (1974) Is attention shared between the ears? Cog. Psychol. 6, 190-215.
- Sinnott, J. M. (1974) Human versus monkey discrimination of the /ba/-/da/ continuum using 3-step comparisons. J. Acoust. Soc. Am. 55, S55(A).
- Stevens, K. N. (1960) Toward a model for speech perception. J. Acoust. Soc. Am. 32, 47-55.
- Stevens, K. N. (1972) The quantal nature of speech: evidence from articulatory data. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw Hill), pp. 51-66.
- Stevens, K. N. (1975) The potential role of property detectors in the perception of consonants. In Auditory Analysis and Perception of Speech, ed. by G. Fant and M. A. A. Tatham. (New York: Academic Press), pp. 303-330.
- Stevens, K. N. and A. House. (1972) Speech perception. In Foundations of Modern Auditory Theory, ed. by J. V. Tobias, vol. 2. (New York: Academic Press), pp. 1-62.
- Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653-659.
- Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper.



- (1970) Motor theory of speech perception: a reply to Lane's critical review. Psychol. Rev. 77, 234-249.
- Studdert-Kennedy, M., D. P. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: evidence from a dichotic study. Cog. Psychol. 2, 455-466.
- Studdert-Kennedy, M., D. P. Shankweiler, and S. Schulman. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Am. 48, 579-594.
- Sumby, W. H. and I. Pollack. (1954) Visual contributions to speech intelligibility in noise. J. Acoust. Soc. Am. 34, 58-60.
- Tallal, P. (in press) Implications of speech perception research to clinical populations. In "Implications of Basic Speech and Language Research for the School and Clinic" conference proceedings, ed. by J. F. Kavanagh and W. Strange. (Cambridge, Mass.: MIT Press).
- Tallal, P. and M. Piercy. (1973) Developmental aphasia: impaired rate of nonverbal processing as a function of sensory modality. Neuropsychologia 11, 389-398.
- Tallal, P. and M. Piercy. (1974) Developmental aphasia: rate of auditory processing and selective impairment of consonant perception. Neuropsychologia 12, 83-93.
- Tallal, P. and M. Piercy. (1975) Developmental aphasia: the perception of brief vowels and extended stop consonants. Neuropsychologia 13, 69-74.
- Tartter, V. C. and P. D. Eimas. (1975) The role of auditory feature detectors in the perception of speech. Percept. Psychophys. 18, 293-298.
- Tulving, E. (1972) Episodic and semantic memory. In Organization of Memory, ed. by E. Tulving and W. Donaldson. (New York: Academic Press), pp. 381-403.
- Whitfield, I. C. (1965) "Edges" in auditory information processing. In Proceed. 23rd Inter. Cong. Physiolog. Sci., (Tokyo), 245-247.
- Whorf, B. L. (1940) Linguistics as an exact science. Technology Review 43: 61-63, 80-83. [In Language, Thought, and Reality, ed. by J. B. Carroll. (Cambridge, Mass.: MIT Press) 1956, pp. 220-232].
- Wood, C. C. (1974) Parallel processing of auditory and phonetic information in speech discrimination. Percept. Psychophys. 15, 501-508.
- Wood, C. C. (1975) Auditory and phonetic levels of processing in speech perception: neurophysiological and information-processing analyses. J. Exp. Psychol.: Human Percept. Perform. 1, 3-20.
- Wood, C. C., W. R. Goff, and R. S. Day. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.

## Outline of Surname Pronunciation Rules for a Reading Machine

Jane H. Gaitenby and Susan Lea Donald\*

### ABSTRACT

Ad hoc rules for pronouncing U. S. surnames, written to supplement the stored vocabulary of the Haskins reading machine for the blind, are outlined. Basic to the rules is a procedure for assigning stress patterns that will be referred to below--in addition to letter-to-sound rules that will not be discussed in detail here. Hand simulation of the rules' operation has produced reasonable phonetic agreement, in the majority of cases, with readings by humans of hundreds of different surnames. (Names made up of an assemblage of morphemes may occasionally be assigned inappropriate stress patterns, but when pronouncing them by relatively simple letter-based rules, this is unavoidable.)

### INTRODUCTION

The reading machine for the blind, designed and developed at Haskins Laboratories, has been described fully in Nye, Hankins, Rand, Mattingly, and Cooper (1973), and elsewhere. Briefly, it entails a computerized "dictionary" memory of 150,000 English words that contains both the word spellings and the corresponding phonetics. Input text-word spellings are thus matched in the dictionary and immediately converted to phonetic strings. Phrasal stress is assigned next; then the succession of phonetic characters is transformed to sound, all by computer program. The listener hears the text output on audio tape in spoken sentences of ordinary English (in synthetic speech). Figure 1 is a flow chart showing the overall text-to-speech process.

The prototype reading system has been demonstrably workable for several years. The intelligibility of its synthetic speech output has been appraised in formal tests against the intelligibility of the same corpora in natural speech, and also with the particular aim of discovering areas of difficulty on the segmental level in the machine speech, (Nye and Gaitenby, 1973, 1974). The results, although generally satisfactory in respect to intelligibility,

---

\*Also, University of Connecticut, Storrs.

Machine will accept input in page form and will recognize OCR-A typefont. Maximum operating rates are 30 documents/min, 200 characters/sec. Output medium, digital magnetic tape. Incorporates on-line correction facility.

Computer program containing stored phonemic transliterations and grammatical categories of more than 150,000 English words. Finds phoneme equivalents of each text word and displays output for editorial checking.

Inserts stress and intonation instructions primarily on the basis of lexical rules. Output can also be checked by an editor.

Computes pitch amplitude and formant frequencies of desired acoustic output on the basis of a system of rules.

Special purpose device designed to generate larynx-like waveform or sibilant noise which is modulated by a system of three parallel formant frequency resonators to create intelligible speech. Speaking rate adjustable within wide limits.

A standard audio frequency tape recorder records synthetic speech on 1/4 inch magnetic tape which is conveyed to the researchers at the University.

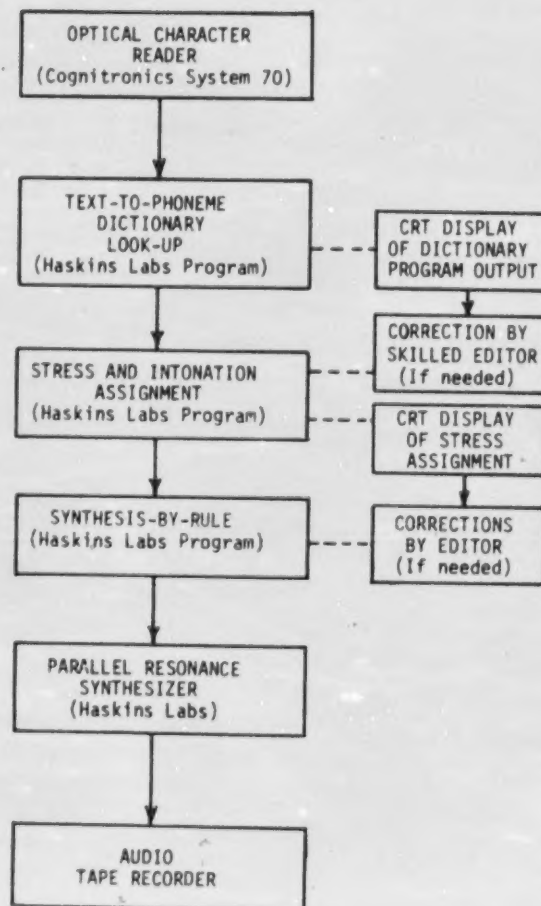


Figure 1: Text-to-speech processor.  
(from Nye, Hankins, Rand, Mattingly, and Cooper, 1973.)

have led to modifications of the synthesis program on several levels, and improvements are still under way. New synthesis programs are being designed that are directed toward the achievement of more natural speech timing and quality; and in another area, plans are afoot to incorporate the algorithm described in this paper within the dictionary search procedure.

At present, words that are not part of the vocabulary stored in the reading machine are provided by a human editor who fills the occasional gaps in the text--as displayed on a computer terminal after the dictionary matching operation--by typing in the needed phonetics for each missing word.

The stored dictionary, large as it is, is not sufficiently supplied with proper names, except for those that are frequent, or those that chance to be stored because they are also everyday words (china, pacific, brown, wood, etc.). Among proper names, surnames stand out as an important class, frequent in texts and frequently absent from the storage. After a preliminary inspection of numerous surnames, in which useful structural regularities were noted, we decided to write letter-based rules for their conversion to phonetics as one move toward enlarging the machine's effective vocabulary.

The rules we have written have been derived empirically, and in an ad hoc fashion. They are intended to convert the prevalent varieties of last names from printed English to recognizable phonetics, but because of the diverse origins of even frequent types of American surnames, errors in pronunciation inevitably occur from time to time when using the rules, as they may when humans attempt to pronounce unfamiliar names. (Some types of French names are especially likely to suffer in the present rules, but fortunately French names are not a prevailing class in English texts). As they stand, the rules demonstrate that customary or at least recognizable pronunciations can be produced for the majority of American surnames, without recourse to actual morph decomposition.

As with all common English or anglicized vocabulary, the pronunciation of names is a function of spelling and stress pattern (not to mention morphology). It is well known that when unstressed, vowels written as single letters, for example, reduce their stressed quality to some extent--and consonantal properties are also affected by stress. Accordingly, a main consideration in constructing the rules has been that of determining ways to assign correct stress patterns to names. (The stress pattern can be defined as the relative prominence of successive syllables in a word; a syllabification procedure is therefore another part of the rules' pronunciation process.)

The rules assign one of three stress values to each syllable: high, mid, or low. These stress specifications are reflected in the speech synthesis program by certain variations in segmental features as well as in one or more of the prosodic parameters: fundamental frequency, duration, and intensity.

Individual last names in the United States tend to be two syllables long, but there are also many thousands of one and three syllable names.



Comparatively few names (fewer thousands!) are four syllables long. (The number of names decreases as the number of syllables contained increases.) We have encountered no names exceeding eight syllables, and only two of those. (One is Venkatasubramanian. The other is given as the next example). One-syllable names present no problem in assigning stress, because, as high information words, they normally receive high stress, in context as well as in isolation. Like many common nouns, the stress of two- and three-syllable name types is most often high on the first syllable and relatively low on the second, followed by mid stress on the last syllable of three-syllable words, unless--as in Anderson, for example--the last morpheme is extremely common, with low stress in that condition. The stress patterns of long names may not be predictable from the mere spelling (for example, Thananjayarajasingham). Nevertheless, certain letters and structural patterns at the end of names often key the stress pattern (except in those cases where the name is a compound that requires morphological analysis, as often happens in very lengthy names, especially some from India or from Russia; for example, Narasimhamurty, Nandeeswaraiya, Blagoveschenskaya).

Thus, in the simpler types of names, the assigned stress pattern is "normal" and unmarked in the pronunciation rules because the pattern depends only on the number of syllables in a name. Special stress pattern assignment also depends on the number of syllables, but only after a marked stress type has been signalled by some particular end-of-name structure. For this reason, the end of each name is inspected at an early step in the operation of the rules.

We have numerically coded the alphabet's 26 letters and the 39 phonemes available in the surname pronunciation rules. The code distinguishes vowels from consonants, indicates stress-dependent vowel alternations, and signifies the place, manner, and voicing of the frequent consonants. The code was phonetically-based in order to try to put the rules into linguistic perspective, although that view has not been exploited to any appreciable degree. (We will refer only briefly to the code itself below, since it involves much detail.)

One work provided significant reference material in constructing the rules: The Structure of English Orthography, by R. L. Venezky. The book deals with spelling and pronunciation regularities for the main vocabulary of English, with a limited consideration of stress effects. Although no attention is given there to proper names, the listings and observations were useful indeed, since normal conventions do tend to be followed in surnames, in spite of the fact that there are some spellings and pronunciations that are peculiar to names.

In applying the rules, we presuppose that capitalized words are noted as such in the reading machine program, and that a search of the dictionary memory is made at the start for each text word in its uncapitalized form. When a word search fails, and the first letter of the word is capitalized, the surname rules come into play. (A number of place names and other proper names will be correctly rendered into phonetics when caught in the surname rule net. Others will not. Only rules based on morph decomposition in

tandem with a morph lexicon storage--like those assembled at the Research Laboratory of Electronics, MIT, for the general vocabulary of English (Hunnicut, 1976)--can ever be expected to cope with compound words, including numerous proper names of all types. A far larger lexicon than even the RLE's current storage of 11,000 English morphs will be required to deal realistically with compound surnames and other compound proper names, in view of the fact that morphs in names represent the entire range of world languages.)

#### OUTLINE OF SURNAME PRONUNCIATION RULES, WITH NOTES

The procedure we use for surname print-to-phonetics conversion is shown in Table 1, and two examples of the conversion process are given in Figure 2. A short elaboration on the steps in the process follows.

Individual letters and their most probable phonetic reflexes are uniquely specified in Step 1, that is, A(a)=[æ], B(b)=[b]. The letters C, Q, and Y, however, are replaced by quasiphonetic coding because their phonetic values can be approximated only when orthographic context is ascertained (in Step 3). The letter X also receives interim coding, there being no single phoneme to replace it. Initial and final space indicators are coded as well as the phonetic characters because they are also participating members of contexts that affect pronunciation. Contexts (both positional and prosodic) are considered in subsequent steps.

A total of 39 phonemes is employed in the rules, with only 22 assignable at the first stage, as has just been explained. The phoneme /ð/ is omitted entirely from the possible final transcriptions due to the fact that /ð/ is very rare in last names, and seems to be adequately suggested in those rare cases by the use of /θ/. This is one difference between the probabilities of surname pronunciation and that of the most common vocabulary. In the latter, due to the overwhelming frequency of "the" and deictic words, /ð/ is a far more probable reflex for the spelling TH than /θ/, although in the much less frequent words of the everyday lexicon, /θ/ is the most likely TH reflex (Allen, 1976). It should be noted that the spelling TH, when preceded by space and followed by OM, is pronounced /t/ in the surname rules, as in Thomas, Thompson, and some 30 other names. This latter reflex for TH probably does not occur in standard pronunciation of the general vocabulary.

The purpose of Step 2 is to modify the pronunciation of vowels at or near the end of a name, or to specify that a change from the normal (unmarked) stress pattern is to be made, or both. A list of 35 common types of name endings, signaling shifts from the first assigned pronunciation, is here used to test the end of a name for a match [for example, -ie(+s), -CVCe, -VCV, -CCV, -CVCVm/n/r/l(-s).] Following this, one of 12 short groups of ordered subrules is invoked if a match occurs. The subrules dictate such changes as diphthongization of penultimate vowel, replacement of final E by silent E, or modification of quality in a final single vowel (other than E) with an accompanying stress tag denoting a marked stress pattern for the word.

TABLE 1: PROCEDURE FOR SURNAME PRINT-TO-PHONETICS CONVERSION

(Scan word; store letters in succession.)

1. Replace each letter by (most probable) phonetic symbol. Add space symbol before the first letter and after the last.
2. Scan right to left. Test the right end of word for a match. If a match occurs, modify phonetics and/or tag for special stress, if called for.
3. Scan left to right. Test for sequence matches, moving successively across the word. If a match is made, replace the character(s) by modified phonetics.
4. Syllabify
5. Count syllables and assign stress pattern (normal pattern, unless tagged otherwise in "2" above): insert stress symbols into phonetic string.
6. Copy phonetic and stress-marked string, replacing vowels by other vowels (where necessary) according to stress value of syllable in which given vowel appears.
7. Store resulting phonetics for the name in correct text location, for synthesis.

FIGURE 2: TWO EXAMPLES OF SURNAME PRINT-TO-PHONETICS CONVERSION

With Unmarked Stress Pattern

Name: | R e c h n i t z |

Step 1 initial space (i.s.) | r e @ h n i t z | final space (f.s.)

2 Significant ending?  
No.  
Copy as in I.

3 i.s. | r e @ h n i t z | f.s.  
Via context: ↓ ↓  
i.s. | r e k n i t s | f.s.

4 Syllabify C V C / C V C C

5 2 syllables, unmarked stress: HL pattern  
Thus, <sup>H</sup>C V C <sup>L</sup> C V C C, or  
6 <sup>H</sup>r e k <sup>L</sup>n i t s ( [I] does not change in Low stress )

7 Store | r e k n i t s . |

With Marked Stress Pattern

Name: | J a l o s k y |

1 i.s. | j æ l a s k y | f.s.

2 Significant ending?  
Yes: | C C y | f.s.  
Change final y to i\*; Stress Tag LHL.  
Copy II as revised.

3 i.s. | j æ l a s k i | f.s. + LHL  
Via context: No change.

4 i.s. | j æ l a s k i | f.s. + LHL  
Syllabify C V / C V C / C V

5 3 syllables, tagged LHL stress pattern.  
<sup>L</sup>C <sup>V</sup>C <sup>H</sup>C <sup>V</sup>C <sup>L</sup>C V, or  
6 <sup>L</sup>j æ l a s k i ( [æ] changes to [ə] in Low stress )

7 Store | j æ l a s k i . |

\* A vowel has permanent color in this case.



Whether or not modified in Step 2, the name is scanned from left to right in Step 3 to permit alteration of the previously established phonetic string on the basis of "letter" context. Here, sequential characters of the word are tested for a match with a list of 209 individual sequences. A match found at this level requires a new replacement operation; that is, the previous phonetic representation of one or more successive letters will be altered. Each new matching attempt across the word involves two or more characters to the immediate right of one that has just been replaced or copied (that is, brought down unchanged from Step 2 because no match was found during Step 3).

To establish and insert syllable boundaries (Step 4), the name is scanned with reference merely to the consonant (C) or vowel (V) status of the string of phonetic characters. Table 2 summarizes the syllabification scheme, (Gaitenby, 1967).

TABLE 2: Syllabification rules.

| If this sequence             | Then boundary (/) is:         |
|------------------------------|-------------------------------|
| (space) CV (space) . . . . . | CV/                           |
| " VC "                       | VC/                           |
| VV                           | V/C                           |
| VCV                          | V/CV                          |
| VCCV                         | VC/CV                         |
| VCCCV                        | VCC/CV                        |
| VCCCCV                       | VCC/CCV                       |
| VCCCCCV                      | VCC/CCCCV (as in "Sandstrom") |

Note: a syllable boundary is not permitted within a string of initial or final consonants.

Step 5 consists of a syllable count followed by a check for the presence of a stress tag indicating a marked stress pattern. (The number of syllables is equal to the number of vowels in the name at this stage.)

In the unmarked case--when a name has acquired no stress tag in Step 2--the stress pattern is directly determined by the number of syllables in the word, as shown in Table 3.

TABLE 3. Unmarked (normal) stress pattern.

| If the number<br>of syllables<br>in the name is | Then the Stress Pattern is: |     |     |                |
|-------------------------------------------------|-----------------------------|-----|-----|----------------|
|                                                 | 1st                         | 2nd | 3rd | 4th Syllable   |
| 1                                               | High                        |     |     | e.g., Hamm     |
| 2                                               | H                           | Low |     | Hammer         |
| 3                                               | H                           | L   | Mid | Hammerschmidt  |
| 4 & 4+                                          | (All High ...)              |     |     | (Hammerbacher) |

(Most names of four syllables, and some of the longer names, will have been tagged for special stress in Step 2 by virtue of their rightmost components. Therefore, the monotonous pattern shown for the "4 & 4+" syllable case in Table 3 above, will seldom be assigned.)

If, on the other hand, the name has been tagged for special stress, it will be assigned by the tag to one of the three marked stress classes shown in Table 4. As before, the number of syllables in the name then automatically determines the specific pattern within the class.

TABLE 4: MARKED STRESS PATTERNS (assigned by reference to end of name structure).

| <u>Number of<br/>Syllables<br/>in Name</u> | <u>Pattern L H L</u> |   |   |     | <u>Pattern H L L</u> |   |     |     | <u>Pattern M L H</u> |   |     |   |
|--------------------------------------------|----------------------|---|---|-----|----------------------|---|-----|-----|----------------------|---|-----|---|
| 2                                          |                      |   |   | H L |                      |   |     |     |                      |   |     |   |
| 3                                          |                      |   | L | " " |                      |   |     |     |                      |   | L   | H |
| 4                                          |                      | M | " | " " |                      |   | H   | L L |                      | M | " " | " |
| 5                                          |                      | L | " | " " |                      |   | L   | " " |                      | L | " " | " |
| 6                                          | H                    | " | " | " " |                      |   | M   | " " |                      | H | " " | " |
|                                            |                      |   |   | " " | L                    | " | " " | " " |                      | L | " " | " |

Examples:

|              |             |            |
|--------------|-------------|------------|
| Brot ta      | -- --       | Chicotte   |
| Las ky       |             | La fosse   |
| A mi co      | Ro me o     | Lachapelle |
| Pu las ki    | Mo vi us    |            |
| Sa wa shi ma | Va le ri o  |            |
| Ma ri nel lo | Au re li us |            |

It can be seen that both the marked and unmarked stress patterns generally employ some successive portion of one underlying stress alternation: high, low, mid, low, ... (Gaitenby, 1967). Only the "anchor" region of each pattern differs from the others. The normal pattern is anchored in "high" stress, at the beginning of a name, but the marked patterns are anchored at the end. The normal pattern, H L M, and the tag pattern, L H L, overlap in the two-syllable name condition, and there produce identical stress patterns: H L. An apparent exception to the basic series of regular stress alternations is found in the final two syllables of the tag pattern H L L. However, that final sequence of two low-stressed syllables usually represents two vowels (for example, i ) that collapse by rule into a single low syllable (for example, y ) in context--thus fitting the general alternation pattern.

Step 6 begins with the phonetic characters of the name in proper sequence, with an assigned stress mark inserted before each phonetic syllable. Each vowel that is a member of a mid or low stressed syllable is now appropriately replaced, if it is replaceable, by reference to a listing of stress-graded phonetic reflexes. There are 40 available patterns of vowel shifts by stress grade.

The final Step, 7, moves the stress-marked, vowel-modified character string into its correct location in the text, to be synthesized with the sentence of which it is a part.

#### SUMMARY

The scheme of the rules is to substitute the most probable single phonetic symbol unique for each letter of the name at the outset, and then to modify the initial string according to the context, both segmental and suprasegmental. Changes from the "normal" stress pattern, indicated at the right end of a word, have global effects on the pronunciation, and are therefore sought out and designated at an early stage in the rules. Other shifts in the phonetics, local in scope, are made at later stages. A name that is not subject to modification in its surface pronunciation, using the rules described, will have achieved its final phonetic form by the end of Step 1.

#### REFERENCES

- Allen, J. (1976) Synthesis of speech from unrestricted text. Proc. IEEE 64, 4, 436.
- Gaitenby, J. H. (1967) Rules for word stress analysis for conversion of print to speech. Haskins Laboratories Status Report on Speech Research SR-12, 131.
- Hunnicut, S. (1976) A new morph lexicon for English. In Preprints for Sixth International Conference on Computational Linguistics (COLING),

Ottawa, Canada, June 28.-July 2, 1976.

- Nye, P. W. and J. H. Gaitenby. (1973) Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test results). Haskins Laboratories Status Report on Speech Research SR-33, 77.
- Nye, P. W. and J. H. Gaitenby. (1974) The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research SR-37/38, 169.
- Nye, P. W., J. D. Hankins, T. C. Rand, I. G. Mattingly, and F. S. Cooper. (1973) A plan for the field evaluation of an automated reading system for the blind. IEEE Trans. Audio Electroacoust. AU-21, 3, 265.
- Venezky, R. (1970) The Structure of English Orthography. (The Hague: Mouton).



## Building an -S Detection and Removal Algorithm

George Sholes

### ABSTRACT

In converting print to synthetic speech, we may take into account the ending -S, which forms the plural of most nouns and the third person singular of the present tense of most verbs. An algorithm for detecting and removing spelled -S from print-words is described and its accuracy in certain cases is estimated.

In scanning a line of printed English to yield an intelligible and speechlike output, a reading machine must cope with the vagaries of English orthography even while taking advantage of its many regularities. That is, an algorithm that derives a string of phonemes from the printed word to serve as input to a program of speech synthesis by rule must yield different vowel symbols for such orthographically similar words as 'gown' and 'grown'. For English, and some other written languages, it is easy to multiply these examples.

A subset of the problems involved in going from spelling to sound involves morphophonemic alternations without accompanying spelling changes. Examples are the first syllable of 'nation/nationality' and the vowel in the second syllable of 'causal/causality'. Attention here is restricted to the handling of one problem in this group, namely the orthographic form of the suffix usually indicating plurality in nouns and the third person singular of the present tense of most verbs. This suffix will be symbolized with -S.

A detailed description of how to spell English plurals is given in Webster's Third New International Dictionary (Gove, 1961). This, of course, includes plurals in -S, that is, those that are pronounced [s], [z], or [əz]. These plurals in -S are spelled with either -s or es. The spelling and the pronunciation of -S go their separate ways. Knowing one gives little immediate information about the other. However, if it can be determined that a print-word does end with spelled -S, if the -S can be removed, and if the print-word minus spelled -S can be pronounced, then the correct selection among the three pronunciations of the -S suffix can be made according to the last phoneme of the pronunciation. If the pronunciation of the base ends in [s z ʃ ʒ c j], then [əz] is selected for the pronunciation of -S. If the pronunciation ends in [p t k f θ], then [s] is selected. Failing these, the pronunciation [z] represents -S, that is, the pronunciation of the base ends in [b d g v ð m n ŋ y w r l] or a vowel (c.f. Bolinger, 1968:58-59). A

---

[HASKINS LABORATORIES: Status Report on Speech Research SR-48 (1976)]

description of an algorithm for pronouncing certain spelled words will be found elsewhere in this Status Report. Thus, it remains for an -S detection and removal algorithm to: 1) determine whether a particular print-word does or does not end with the spelled -S suffix, and 2) remove the spelled -S suffix if present. The word can now theoretically be pronounced by algorithm. If the -S suffix was present in the original spelling, the correct pronunciation for it can be added on according to the rule given above.

By a happy coincidence the same -S suffix that forms noun plurals is also used to form the regular third person singular of English verbs. The one -S algorithm does a considerable amount of work.

The spelling of the -S suffix is either -s or -es. However, before the spelling is added onto some base print-word, it may be necessary to adjust the spelling of the base. The most familiar such change is given by the rule: y to i and add -es. 'Lady' changes to 'ladies', and 'courtesy' changes to 'courtesies'. Proceeding in the opposite direction, the algorithm removes the -S ending and restores the proper base spelled form: take off -es and change i back to y.

A fundamental ambiguity is that certain English words are spelled in such a way that is impossible, save with lists, to determine whether an -S suffix is present and if it is, how much of the spelling constitutes the -S suffix. For instance, which word has an -S suffix: 'pianos' or 'rhinoceros'? How much of the words 'overshoes' and 'embargoes' is the suffix and how much is the base?

In such circumstances, an -S detection and removal algorithm can only do the least damage, make the fewest mistakes. The number of English print-words in plain noun or verb form that end with the spelling -os, such as 'rhinoceros', may be smaller than the number of those ending with the spelling -o, which form their -S form by adding the spelling -s. One would expect the list of words with 'monopteros' to be much shorter than the list of words with 'innuendo'. With this information, one could formulate a part of the algorithm as follows: If a print-word ends in spelled -os, the -S suffix is present and is removed by deleting the terminal spelling -s.

In order to build an -S detection and removal algorithm it is important to have various compared lists of print-words, where membership in the lists is determined by the way the spellings of the included words end. Comparison of such lists, especially of their total lengths, can suggest which list is best taken as exemplifying the rule and which list the exceptions. A list of 66,500 English base-forms in reversed spelling is available in the English Word Speculum, vol 3, compiled by J. L. Dolby and H. L. Resnikoff (1967).

The algorithm is designed for reading machines and will be used only when consultation of the reading machine pronouncing dictionary fails to show a print-word. This pronouncing dictionary contains both plain forms (bases) and forms with suffix -S. So in certain cases it would be possible, at least in principle, to force a part of the algorithm to work perfectly by entering an exhaustive list of exceptions into the pronouncing dictionary. This will work insofar as a subportion of an open-class word list can be fairly termed

exhaustive.

Here is an illustration of how lists from the English Word Speculum guided construction of an algorithm. The list of words ending in spelled -o, such as 'piano', is 803 members long. The spelling -o for a base implies an -S form in either -s or -es, sometimes indifferently, for example 'cargos', and 'cargoes.' A rule for recovering the plain-form from words ending in -os and -oes would be: take off the -s or the -es respectively. It can be seen that this rule can fail in two different ways. It can take off an -S when none is present, for example, 'rhinocero/s', and it can take off the wrong -S ending, for example, 'oversho/es'. There are 63 base-forms ending in -os and 2 ending in -oes, all of which could have the -S inadvertently removed. There are 34 base-forms ending in -oe, and from these the -S will be removed incorrectly. Eight out of nine print words ending in -os and -oes will be analyzed correctly and one out of nine will not. In the case of base-forms ending in -os and -oes, the right list is the rule we have made.

In using base-forms without further qualification, we are making estimates rather than exact counts. Only words used as verbs or nouns regularly take an -S ending. Adjectives, adverbs and structural words do not. Yet verbs and nouns so far outnumber the others that a list of all spelled words can serve as an approximation to a list of the verbs and nouns.

The rule for -S spelled -s is fairly straightforward. Final spelled -s preceded by a consonant other than -s is -S, for example, 'desks' and 'lamps'. Final spelled -s preceded by another -s is no ending at all, for example, 'business.' Preceded by a, o, or y, final -s is also -S, for example, 'cameras', 'cargos', and 'Marys'. Preceded by i or u, final -s is no ending at all, for example, 'circus' and 'tennis'. For consonants, this rule is believed to be close to perfect. The rule for vowels is estimated by Speculum list lengths to be right 95 percent of the time for -a/s, 73 percent of the time for -is, 93 percent for -o/s, and 94 percent for -us. Examples of errors are 'canvas' and 'pathos', which will have -s incorrectly removed, and 'taxi/s' and 'bureau/s', which will fail to have the -S removed when it should be. In summary:

From -is, -us, -ss, remove nothing, as no -S is present. From -es, see rule which follows. From all other Cs or Vs, -S is present and is removed by taking off -s (including -as, -os, -ys).

The rule for -S spelled -es is more complicated. The rule for adding -es follows: basically, words ending in -s, -z, -x, -ch, and -sh regularly add -es for -S; words ending in other consonants regularly add -s. The problem arises in removing -S from base-forms ending with -e: for example, 'amuse', one of the five spellings that regularly take -es instead of -s. When this happens, there is no way to tell whether -ses represents -s es or -se s, for example, 'buses' and 'amuses'. In this particular case, the rule is to take off -es, as words ending in -s are more numerous than words ending in -se. This rule will be right 82 percent of the time.

The rule for words ending in -xes, -ches, and -shes is the same as for words ending in -ses, that is, -S is present and is removed by taking off final -es. The rule will be correct for -xes ninety-seven percent, for -ches

eighty-three percent, for -shes one-hundred percent, and for -ses eight-two percent of the time. The rule for words ending in -zes follows: -S is present and is removed by taking off final -s. This will be correct 95 percent of the time. The general rule for other consonants ending in -Ces is: -S is present and is removed by taking off final -s.

The rules for vowels differ according to the vowel: for -aes, -ees, -ues and -yes, -S is present and removed by taking off final -s. For -oes, and -ies, -S is present and is removed by taking off final -es. For -ies, i in final position after removal of -es is changed to y. In summary:

Take off -es from the following word endings: -s/es, -x/es, -ch/es, -sh/es, -o/es, -i/es (and change i to y).

From all other -es endings take off -s (including -ze/s, -ae/s, -ue/s, -ee/s, -ye/s, and all other consonants -e/s).

#### SUMMARY

The algorithm built with these compared lists assumes that all spellings ending in -s may contain the suffix -S. All changes depend on what precedes the -s in the word. In the summary form below, a slash marks off the letters to be removed for -S. No slash means no -S is present. The order of the sequence is important, as the last term must come last:

-is, -us, -ss, -s/es, -x/es, -ch/es, -sh/es, -o/es, -i/es, and y i, -/s.

The procedure outlined here goes a long way toward the correct separation and phonetic interpretation of the suffix -S. Much work remains to be done on the detection and interpretation of the orthographic reflexes of other kinds of morphophonemic complexity in English.

#### REFERENCES

- Gove, P. B. (1961) Webster's Third New International Dictionary of the English Language Unabridged. (Springfield: Merriam).  
 Dolby, J. L. and H. L. Resnikoff. (1967) The English Word Speculum, Vol. 3, The Reverse Word List. (The Hague: Mouton).  
 Bolinger, D. (1968) Aspects of Language. (New York: Harcourt, Brace & World).



II. PUBLICATIONS AND REPORTS

III. APPENDIX

PUBLICATIONS AND REPORTS

- Abramson, Arthur S. (1976) Tai tones as a reference system. In Tai Linguistics in Honor of Fang-Kuei Li, ed. by Thomas W. Gething, Jimmy G. Harris, and Pranee Kullavanijaya. (Bangkok: Chulalongkorn University Press) pp. 1-12.
- Bell-Berti, Fredericka. (1976) An electromyographic study of velopharyngeal function in speech. Journal of Speech and Hearing Research, 19, 225-240.
- Borden, Gloria J., M. F. Dorman, F. J. Freeman, and L. J. Raphael. (1976) Electromyographic changes with delayed auditory feedback of speech. Journal of Phonetics, 4.
- Cutting, James E. and B. S. Rosner. (1976) Discrimination functions predicted from categories in speech and music. Perception and Psychophysics, 20, 87-88.
- Hadding, K., H. Hirose, and K. S. Harris. (1976) Facial muscle activity in the production of Swedish vowels: an electromyographic study. Journal of Phonetics, vol. 4, 233-246.
- Nye, Patrick W. (1976) Reading devices for blind people. Medical Progress Through Technology 4, 11-25.

# APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers SR-21/22 to SR-45/46:

| Status Report                   | DDC        | ERIC       |
|---------------------------------|------------|------------|
| SR-21/22 January - June 1970    | AD 719382  | ED-044-679 |
| SR-23 July - September 1970     | AD 723586  | ED-052-654 |
| SR-24 October - December 1970   | AD 727616  | ED-052-653 |
| SR-25/26 January - June 1971    | AD 730013  | ED-056-560 |
| SR-27 July - September 1971     | AD 749339  | ED-071-533 |
| SR-28 October - December 1971   | AD 742140  | ED-061-837 |
| SR-29/30 January - June 1972    | AD 750001  | ED-071-484 |
| SR-31/32 July - December 1972   | AD 757954  | ED-077-285 |
| SR-33 January - March 1973      | AD 762373  | ED-081-263 |
| SR-34 April - June 1973         | AD 766178  | ED-081-295 |
| SR-35/36 July - December 1973   | AD 774799  | ED-094-444 |
| SR-37/38 January - June 1974    | AD 783548  | ED-094-445 |
| SR-39/40 July - December 1974   | AD A007342 | ED-102-633 |
| SR-41 January - March 1975      | AD A103325 | ED-109-722 |
| SR-42/43 April - September 1975 | AD A018369 | ED-117-770 |
| SR-44 October - December 1975   | AD A023059 | ED-119-273 |
| SR-45/46 January - June 1976    | AD A026196 | ED-123-678 |
| SR-47 July 1 - September 1976   |            | ED-128-870 |

AD numbers may be ordered from: U.S. Department of Commerce  
National Technical Information Service  
5285 Port Royal Road  
Springfield, Virginia 22151

ED numbers may be ordered from: ERIC Document Reproduction Service  
Computer Microfilm International Corp. (CMIC)  
P.O. Box 190  
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.

UNCLASSIFIED

Security Classification

## DOCUMENT CONTROL DATA - R &amp; D

Security Classification of title, body of abstract and indexing annotation must be entered when the overall report is classified.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |  |                                                                             |                  |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|-----------------------------------------------------------------------------|------------------|
| 1. ORIGINATING ACTIVITY (Corporate author)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |  | 2a. REPORT SECURITY CLASSIFICATION                                          |                  |
| Haskins Laboratories, Inc.<br>270 Crown Street<br>New Haven, Connecticut 06510                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |  | Unclassified                                                                |                  |
| 3. REPORT TITLE                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |  | 2b. GROUP                                                                   |                  |
| Haskins Laboratories Status Report on Speech Research, No. 48, October - December 1976                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |  | N/A                                                                         |                  |
| 4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |  |                                                                             |                  |
| Interim Scientific Report                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |  |                                                                             |                  |
| 5. AUTHOR(S) (First name, middle initial, last name)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |  |                                                                             |                  |
| Staff of Haskins Laboratories; Alvin M. Liberman, P.I.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |  |                                                                             |                  |
| 6. REPORT DATE                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |  | 7a. TOTAL NO. OF PAGES                                                      | 7b. NO. OF PAGES |
| December 1976                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |  | 345                                                                         | 678              |
| 8a. CONTRACT OR GRANT NO.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |  | 9a. ORIGINATOR'S REPORT NUMBER(S)                                           |                  |
| DE-01774 RR-5596<br>HD-01994<br>V101(134)P-342<br>N00014-76-C-0591<br>DAAB03-75-C-0419(L433)<br>N01-HD-1-2420                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |  | SR-48 (1976)                                                                |                  |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |  | 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) |                  |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |  | None                                                                        |                  |
| 10. DISTRIBUTION STATEMENT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |  |                                                                             |                  |
| Distribution of this document is unlimited.*                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |  |                                                                             |                  |
| 11. SUPPLEMENTARY NOTES                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |  | 12. SPONSORING MILITARY ACTIVITY                                            |                  |
| N/A                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |  | See No. 8                                                                   |                  |
| 13. ABSTRACT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |  |                                                                             |                  |
| <p>This report (1 October - 31 December 1976) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation of its investigation, and practical applications. Manuscripts cover the following topics:</p> <ul style="list-style-type: none"> <li>-Issues in Theory of Action</li> <li>-Physiological Aspects Speech Production: Why Study Speech Production?</li> <li>-Universals in Phonetic Structure, Role in Linguistic Communication</li> <li>-Difference Limens for Formant-Frequencies for Steady-State and Consonant-Bounded Vowels</li> <li>-Vocal Tract Normalization for /s/ and /š/</li> <li>-Speech, the Alphabet and Teaching to Read</li> <li>-Visual Processing and Short-term Memory</li> <li>-Contrasting Orientations to Theory of Visual Information Processing</li> <li>-Evidence for Special Speech Perception Subsystem in Human</li> <li>-Further Observations on Role of Silence in Perception of Stop Consonants</li> <li>-Perception of Implosive Transitions in VCV Utterances</li> <li>-What Can /w/, /l/, /y/ Tell Us About Categorical Perception?</li> <li>-Laterality and Localization: Right-ear Advantage for Speech Heard on Left</li> <li>-Left-ear Advantage for Sounds characterized by Rapidly Varying Resonance Frequency</li> <li>-An Information-Processing Approach to Speech Perception</li> <li>-Outline of a Surname Pronunciation--Rules for Reading Machine</li> <li>-Building an -S Detection and Removal Algorithm</li> </ul> |  |                                                                             |                  |

DD FORM 1473

(PAGE 1)

UNCLASSIFIED

S/N 0101-807-6811

\*This document contains no information not freely available to the general public.  
It is distributed primarily for library use.

Security Classification

A-31408

360



UNCLASSIFIED

Security Classification

| 14<br>KEY WORDS                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | LINK A |    | LINK B |    | LINK C |    |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|----|--------|----|--------|----|
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | ROLE   | WT | ROLE   | WT | ROLE   | WT |
| Action Theory - Issues<br>Speech Production - Physiological Aspects<br>Phonetic Structure - Role in Linguistic Communication<br>Formant Frequencies - Difference Limens for Vowels<br>Vocal Tract Normalization - /s/ and /š/<br>Speech, Alphabet, Reading<br>Visual Processing - Memory<br>Visual Information Processing - Contrasting Orientations<br>Speech Perception - Evidence for Human Subsystem<br>Stop Consonants - Role of Silence, Perception<br>VCV Utterances - Perception<br>Categorical Perception - /w/, /l/, /y/<br>Ear Advantage - Speech, Right<br>Ear Advantage - Rapid, Frequency Changes, Left<br>Speech Perception - Information-Processing<br>Surname Pronunciation - Rules<br>S Detection - Algorithm for Text |        |    |        |    |        |    |

361

DD FORM 1473 (BACK)

S/N 0101-407-6021

UNCLASSIFIED

Security Classification

A-31409

# ERRATA -- SR/48

Status Report 48 was an extremely lengthy issue produced in a tight schedule; this factor plus the changeover from the manual to the computer processing system resulted in a number of errors in the report.

Here follows an Errata Sheet to supplement SR48:

Perception of Implosive Transitions in VCV Utterances  
Bruno H. Repp.

pp. 209-233

p. 214, par. 2, line 6: denoted here by F should be denoted here by  $\hat{F}$ .  
p. 216, par. 6, line 7: # should be  $\neq$   
p. 216, par. 7, line 2: < should be >  
p. 222, par. 7, line 2:  $VC^1-C_1V$  should be  $VC_1-C_1V$ .  
following p. 228: TABLE 1

TABLE 1: Rough estimation of VCV reaction times (RTs) as a mixture of VC and CV reaction times.

|                                                   | Closure Duration |      |      |      |      |
|---------------------------------------------------|------------------|------|------|------|------|
|                                                   | 65               | 90   | 115  | 140  | 165  |
| Percent correct, p(C)                             | 65.3             | 72.2 | 79.9 | 77.8 | 88.2 |
| Percent errors, p(E)                              | 34.7             | 27.8 | 20.1 | 22.2 | 11.8 |
| Estimated p(VC)                                   | 30.6             | 44.4 | 59.8 | 55.6 | 76.4 |
| Estimated p(CV)                                   | 69.4             | 55.6 | 40.2 | 44.4 | 23.6 |
| Actual V-CV RTs, $\bar{t}(V-CV)$                  | 501              | 521  | 578  | 555  | 616  |
| Actual VCV RTs, $\bar{t}(VCV)$                    | 444              | 466  | 432  | 457  | 468  |
| Estimated VCV RTs, $\hat{t}(VCV)$                 | 469              | 465  | 469  | 466  | 447  |
| Observed-estimated, $\bar{t}(VCV) - \hat{t}(VCV)$ | -25              | 1    | -37  | -9   | 21   |

An Information-Processing Approach to Speech Perception  
James Cutting and David Pisoni

p. 287-325

- p. 289, par. 4, line 1: should be: evidence, more compatible  
p. 292, par. 2, line 14: should be: and may in a pneumatic  
p. 293, par. 3, last line: hetarchical should be heterarchical  
p. 298, par. 4, line 4: time-lagged should be time-tagged  
p. 303, par. 4, line 15: right pane should be right panel  
p. 305, par. 4, line 18: redistinctive should be re distinctive  
p. 309, par. 3, line 8: /ntak/ should be /ntək/  
p. 310, par. 2, lines 7-8: Studdert-Kennedy, Shankweiler and Schulman,  
1970, should be Studdert-Kennedy, Liberman,  
Harris and Cooper, 1970.

Outline of a Surname Pronunciation--Rules for a Reading Machine

Jane Gaitenby and S. Lea Donald

p. 327-337

- p. 331, par. 3, line 2: [æ] should be [æ]  
p. 334, Table 2, line 3, right: V/C should be V/V  
p. 334, Table 2, line 8, right: VCC/CCCCV/ should be VCC/CCCV  
p. 336, par 1, line 11: (for example, i) should be (for example, iə)  
p. 336, par. 1, line 12: (for example, y) should be (for example, yə)

Universals in Phonetic Structure and Their Role in Linguistic Communication

Michael Studdert-Kennedy

p. 43-50

- p. 43, par 1, line 3: perceptual should be syllable  
p. 43, par 2, line 14: will should be with  
p. 44, par. 2, line 2: (Harris, in press) should be (Hockett, 1958)  
p. 46, par. 5, line 2: (Hockett, 1958) should be (Lane et al., 1976)  
p. 46, par. 5, line 10: (Bellugi and Klima, 1975) should be  
(Bellugi, Klima and Siple, 1975)